UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

Ciências
ULisboa

SPATIO-TEMPORAL METHODS AND MODELS FOR UNEMPLOYMENT ESTIMATION

Doutoramento em Estatística e Investigação Operacional
Probabilidade e Estatística

Soraia Alexandra Gonçalves Pereira

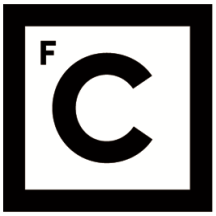
Tese orientada por:
Prof. Doutor Kamil Feridun Turkman e
Dr. Luís Paulo Fernandes Correia

Documento especialmente elaborado para a obtenção do grau de doutor

2018

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS



Ciências
ULisboa

SPATIO-TEMPORAL METHODS AND MODELS FOR UNEMPLOYMENT ESTIMATION

Doutoramento em Estatística e Investigação Operacional
Probabilidade e Estatística

Soraia Alexandra Gonçalves Pereira

Tese orientada por:
Prof. Doutor Kamil Feridun Turkman e
Dr. Luís Paulo Fernandes Correia

Júri:
Presidente:

- Doutor Luís Eduardo Neves Correia, Professor Catedrático da Faculdade de Ciências da Universidade de Lisboa

Vogais:

- Doutor Håvard Rue, Professor
  Department of Mathematical Sciences at Norwegian University of Science and Technology
- Doutora Paula Cristina Sequeira Pereira, Professora Adjunta
  Escola Superior de Tecnologia de Setúbal do Instituto Politécnico de Setúbal
- Doutora Sónia Manuela Ferreira Leite Soutelo Torres, Diretora
  Serviço de Estatísticas do Mercado de Trabalho do Departamento de Estatísticas Demográficas e Sociais do Instituto Nacional de Estatística
- Doutora Maria Manuela Costa Neves Figueiredo, Professora Catedrática
  Instituto Superior de Agronomia da Universidade de Lisboa
- Doutor Kamil Feridun Turkman, Professor Catedrático
  Faculdade de Ciências da Universidade de Lisboa (orientador)
- Doutora Patrícia Cortés de Zea Bermudez, Professora Auxiliar
  Faculdade de Ciências da Universidade de Lisboa

Documento especialmente elaborado para a obtenção do grau de doutor

Fundação para a Ciência e Tecnologia no âmbito da bolsa de doutoramento SFRH/BD/92728/2013

2018

Acknowledgements

First of all, I would like to thank my supervisors: Professor Feridun Turkman and Luís Correia for their invaluable time and encouragement.

Professor Turkman was always very positive, interested and supportive. He transmits confidence and motivation in every meeting and it was a pleasure to work with him.

Luís Correia was the person at the Portuguese National Statistics Institute (INE) who first introduced me to the problem. This would later become the primary focus and motivation of our work.

I would also like to express my gratitude to Professor Antónia Turkman for all the times she helped me with questions about Bayesian approaches and computational problems. Her assistance was greatly appreciated.

My thanks also go to Professor Håvard Rue and his team for the excellent discussion forum about R-INLA and for the access they gave me to their server. These tools were incredibly helpful.

The data used were provided by Paula Marques from INE whom I would like to thank for responding to my requests. Also, the sympathy of my colleagues at the INE office was very important to me and I thank them for their kind and warm hospitality.

My colleagues in the office at the Faculty of Sciences were also very kind and helpful during this time, especially over the last few months, and in particular, I would like to thank Carolina for all her support.

I would also like to thank my boyfriend, family and friends. Luciano was my most precious supporter. My parents and brother are always with me and taught me the most important values I know. I love them as they are! My friends are the best friends in the world.

My thanks also go to my reception institutions, Centro de Estatística e Aplicações (CEAUL) and INE, along with Fundação para a Ciência e Tecnologia (FCT) who gave me financial support through a PhD grant with the reference SFRH/BD/92728/2013.

Thank you also to my English teacher, Henry, for his assistance in reviewing the thesis.

Abstract

In Portugal, the National Statistical Institute (NSI) publishes official quarterly estimates of the labour market for the national territory and for NUTS I and NUTS II regions. NUTS is the nomenclature of territorial units for statistics, communly used by Eurostat, and has three different levels: NUTS I, NUTS II and NUTS III, depending on the disagregated level. The estimation is based on a direct method, using the data from the Labour Force Survey (LFS). However, for NUTS III regions, the sample size of the LFS is not enough to provide accurate estimates using this direct method. This problem is known as the small area estimation problem and it can arise in several disparate areas such as epidemiology, ecology, economics, social sciences, among others.

Within the small area estimation (SAE) framework, several methods and models are suggested and they are centered around the basic Fay-Heriot model and its extensions in several directions. However, the assumptions made in these models are very restrictive and do not appear to be suitable in the context of unemployment.

In this study we propose three alternative approaches for unemployment estimation in small areas.

The first approach is based on generalized linear models (GLM) at areal level, where three different data modelling strategies are considered and compared: modelling of the total unemployed through Poisson, Binomial and Negative Binomial models; modelling of rates using a Beta model; and modelling of the three states of the labour market (employed, unemployed and inactive) by a Multinomial model.

The second approach is based on spatial point processes. From the 4th quarter of 2014 onwards, all the sampling units of the LFS are georeferenced, mainly the residential buildings. For that reason, we propose using this new data, together with the information specific to the families to model the intensity of points and the marks associated to those points, through a marked log Gaussian Cox processes model. Here, the points are the residential buildings, whereas the associated marks are the number of unemployed people residing in these buildings. The basic assumption behind this model is that, although we know the geo-referenced positions of the residential units in the labour sample survey, we do not know the spatial configuration of all residnetial units in the population and therefore, we take the sampled residential units as a realization of a spatialpoint process.

Recently, the NSI provided us with information about the locations of all residential buildings in the national territory. Consequently, it is no longer necessary to model the points, as all the residential buildings are georeferenced. Thus, the third method we propose is based on a point referenced data model, also described as a geostatistical model. This model assumes that the points in the population are

fixed and the interest is to model the spatial variation of the marks. The modelling process is based on a spatial extrapolation of the unemployment figures from the 14000 residential buildings sampled in the LFS to all other known residential units not sampled by the labour survey.

A comparison between the mentioned models, the direct method and the traditional small area models, shows that the geostatistical model is the most favorable due to the good behaviour in terms of variability and the detailed information it can provide.

We follow a Bayesian approach and the inference is made using the package R-INLA in the software R.

Keywords: Unemployment estimation; areal level models; marked spatial point processes; geostatistical model; INLA; small area estimation.

Resumo

Em Portugal, o Instituto Nacional de Estatística (INE) publica trimestralmente as estimativas oficiais do mercado de trabalho a nível nacional e para as regiões NUTS I e NUTS II. NUTS é a nomenclatura das unidades territoriais usada para fins estatísticos, e engloba três níveis hierárquicos: NUTS I, NUTS II e NUTS III, consoante o nível de desagregação. O processo de estimação baseia-se num método direto, usando os dados do Inquérito ao Emprego (IE). Para as regiões NUTS III, a dimensão da amostra do IE não é suficiente para fornecer estimativas precisas usando o método direto. Este é um problema conhecido na literatura como problema de estimação em pequenos domínios e pode surgir em diferentes áreas tal como epidemiologia, ecologia, economia, ciências sociais, entre outras.

Na literatura, têm sido propostos métodos alternativos ao método direto para resolver este problema. O método mais comum é o método Fay-Herriot, um modelo nível área. Contudo, as suposições impostas por este modelo são muito restritivas e não parecem ser adequadas no contexto do desemprego.

Neste trabalho propomos três abordagens alternativas para a estimação do desemprego em pequenos domínios.

A primeira abordagem baseia-se em modelos de regressão nível área, onde são consideradas três estratégias de modelação: modelação do total de desempregados com base em modelos de Poisson, Binomial e Binomial-Negativa; modelação de taxas usando um modelo Beta; e modelação dos três estados do mercado de trabalho (empregado, desempregado e inativo) usando um modelo Multinomial.

A segunda abordagem baseia-se em processos pontuais espaciais. A partir do 4º trimestre de 2014, todas as unidades amostrais do IE foram georreferenciadas, nomeadamente os edifícios residenciais. Por este motivo, propomos usar esta informação bem como informação específica das famílias para modelar a intensidade dos pontos e as marcas associadas a estes pontos, através de um modelo de Cox log Gaussiano. A suposição básica por de trás deste modelo é que, apesar de as localizações dos edifícios residenciais na amostra do IE serem conhecidas, a configuração espacial de todos os edifícios residenciais na população não é conhecida e, portanto, as unidades residenciais amostrais são tratadas como uma realização do processo pontual espacial.

Recentemente, o INE disponibilizou informação sobre as localizações de todos os edifícios residenciais em todo o território nacional. Desta forma, não é necessário modelar os pontos uma vez que agora estes são conhecidos. O terceiro método que propomos baseia-se num modelo de dados referenciados por pontos, ou também conhecido como modelo de geoestatística. Este modelo assume que os pontos na população são fixos e o interesse é a modelação da variação espacial das marcas.

Esta modelação é feita com base numa extrapolação espacial do número total de desempregados a partir dos 14000 edifícios residenciais da amostra do IE para todos os edifícios residenciais que não pertencem à amostra.

A comparação entre os modelos propostos, o método direto e os modelos de estimação em pequenos domínios (SAE) tradicionais mostra que o modelo de geoestatística é o modelo preferencial dado o comportamento em termos de variabilidade e informação detalhada que este pode fornecer.

Neste estudo, seguimos uma abordagem Bayesiana e a inferência foi feita usando o package R-INLA do software R.

Palavras-chave: Estimação do desemprego; modelos nível área; processos pontuais espaciais marcados; modelo de geoestatística; INLA; estimação em pequenos domínios.

Contents

Acronyms

AR Random walk

BLUP Best linear unbiased predictor

CPO Conditional predictive ordinate

CVs Coefficients of variation

DIC Deviance information criterion

EB Empirical Bayes

EBLUP Empirical best linear unbiased predictor

FH Fay-Herriot

GF Gaussian field

GMRF Gaussian Markov random field

GREG General regression estimator

HB Hierarchical Bayes

HT Horvitz-Thompson

iCAR Intrinsic conditional autoregressive

INLA Integrated nested Laplace approximation

LFS Labour Force Survey

LGCP Log Gaussian Cox processes

LGM Latent Gaussian models

MCMC Markov chain Monte Carlo

MH Metropolis-Hastings

NDR National dwellings register

NSI National Statistical Institute

PC Penalised Complexity

PIT Probability integral transform

REML Restricted maximum likelihood

SAE Small area estimation

SPDE Stochastic partial differential equations

WAIC Watanabe-Akaike information criterion

Chapter 1

Introduction

1.1 The problem: spatial distribution of unemployment

In Portugal, the National Statistical Institute (NSI) is responsible for conducting, on a quarterly basis, the Labour Force Survey (LFS) covering the entire national territory, and for disseminating the results to the relevant national and European bodies. Consequently, the NSI publishes official quarterly labour market statistics, including the estimated unemployment figures at different spatial resolutions, typically for NUTS I and NUTS II regions. NUTS is the classification of territorial units for statistics, created by the Eurostat and National Statistical Institutes of the European Union, and includes three hierarchical levels: NUTS I, NUTS II and NUTS III (see figure 1.1 below, and figure A.1 in appendix).

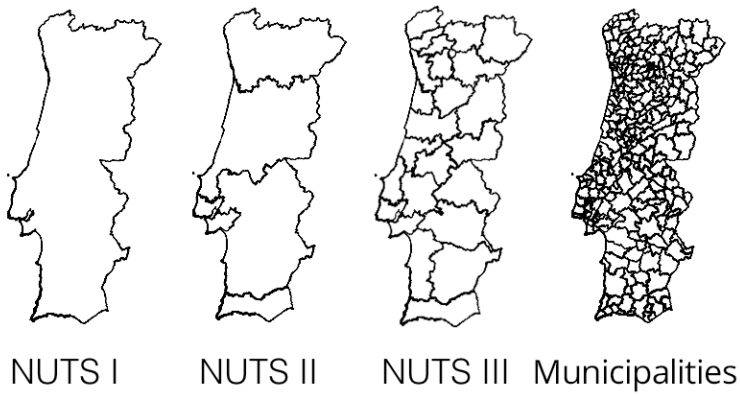


Figure 1.1: NUTS (version 2013) and municipalities in mainland Portugal

Together with the increase in demand for more detailed information at higher spatial resolutions, so too as the demand for more reliable estimates, without incurring the costs associated with larger samples. Typically, the NSI produces unemployment estimates using direct estimation methods based on the Horvitz-Thompson estimator (Horvitz and Thompson, 1952). However, these direct estimation methods do not perform well in smaller geographical areas, meaning that either larger

sample sizes are required, or that small area estimation methods (Rao, 2003), which instead of that borrow strength from neighboring observations must be used.

Considerable methodological developments have evolved to help solve small area estimation problems. The majority of small area methods are based on linear models applied to areal data. These methods can 'borrow strength' from area to area and make use of auxiliary information at a regional level, compensating for the small sample sizes in each area due to the designed sampling survey.

The most frequently used small area methods are based on the basic Fay-Herriot (FH) model, which assumes normal distribution for the direct estimates of the quantity of interest. However, this assumption is not applicable in the context of unemployment, where we intend to model counts or proportions. As a result, we propose some alternative methods to solve this problem. These methods model the observations rather than the direct estimates, contrary to the FH models.

In chapter 2 we suggest using generalized linear models in a hierarchical bayesian approach. We use three different modelling strategies: modelling of the total unemployed through Poisson, Binomial and Negative Binomial models; modelling of rates using a Beta model; and modelling of the three status of the labour market (employed, unemployed and inactive) by a Multinomial model. The implementation of these models is based on the *Integrated Nested Laplace Approximation* (INLA) approach, except for the Multinomial model which is implemented according to the *Markov Chain Monte Carlo* (MCMC) methods.

In chapter 3 we propose an alternative perspective to look at unemployment data through spatial point processes. From 2014 onwards, all the sampling units in the LFS are georeferenced, or rather the dwellings in which the observation units (i.e. individuals) are interviewed. This approach allows for the representation of the sample survey as a realization of a spatial point process together with the associated marks, namely the number of unemployed people in each residential unit. For modelling the intensity of residential unit locations and their associated marks, we suggest a marked Log Gaussian Cox processes (LGCP) model.

Chapter 4 concerns geostatistical modelling. In recent years, the precise locations of all residential buildings in the entire national territory became available. With this new detailed georeferencing information, spatial distribution of residential units is no longer random. Therefore, new spatial models, no longer required to model the randomness of points should, in principle, produce more precise estimates with reduced sampling variation. For this reason, in chapter 4 the objective becomes to model the spatial variation of the marks using point referenced methods, ie the number of unemployed in each sampled residential unit, and then to extrapolate this in space to all georeferenced residential units. In addition to this extrapolation in space, we also extrapolate the results temporally. The temporal extension is based on 9 sequentially observed quarterly sampling surveys (from the 4th quarter of 2014 to the 4th quarter of 2016). For the modelling process, we suggest a geostatistical model with a temporally and spatially structured random effect.

A comparison between the proposed methods, the direct method, and the traditional SAE methods is made in chapter 5. The R-code and the programs used are described in chapter 6.

Chapters 2, 3 and 4 of the thesis are based on the three papers written during

the phD:

1. Pereira, S., Turkman, F., Correia, L. (2016). Spatio-temporal analysis of regional unemployment rates: A comparison of model based approaches. *Revstat.*
   https://www.ine.pt/revstat/pdf/SPATIO-TEMPORALANALYSISOFREGIONALUNEMPLOYMENTRATES.pdf

2. Pereira, S., Turkman, F., Correia, L., Rue, H. (2017a) Unemployment estimation: Spatial point referenced methods and models. *(Submitted to JRSS)*
   https://arxiv.org/pdf/1706.08320.pdf

3. Pereira, S., Turkman, F., Correia, L., Rue, H. (2017b) Spatio-temporal models for georreferenced unemployment data. *(Under preparation)*

Before describing the methodologies we propose, we will first discuss the sampling design of the Portuguese Labour Force Survey (LFS) and the direct estimation method in section 1.2. In section 1.3 we make a review of small area estimation methods. Section 1.4 provides an introduction to the Bayesian approach and section 1.5 contains a summary of both areal methods and point-referenced methods.

1.2 Portuguese Labour Force Survey

1.2.1 Sampling design

The methodologies proposed in this study are highly dependent on the sampling design of the labour force survey (LFS). Therefore, it is important to understand both how the sampling units are drawn and how the inclusion probabilities are calculated.

The LFS is a continuous survey of the population living in private dwellings within the portuguese national territory. The survey provides an understanding of the socioeconomic situation of these individuals during the week prior to the interview (reference week). The dwellings are the sampling units and the inhabitants living in these dwellings are the observation units.

The unemployment figures are published quarterly by the National Statistical Institute (NSI) at both the national and NUTS II level. From one quarter to another, 1/6 of the sample (rotation group) is replaced by a new one. This process is also known as a *rotation scheme*. In this way, each individual in the sample is surveyed over 6 consecutive quarters, inducing strong temporal dependence between the quarterly surveys.

The LFS follows a stratified multi-stage sampling design. First, the sampling frame (National Dwellings Register, built from the 2011 census) was stratified into 25 regions (NUTS III or groups of NUTS III). Then, in each strata, a multi-stage sampling was conducted, where the primary sampling units are areas consisting of one or more contiguous cells of the km^2 INSPIRE grid, and the secondary units are private dwellings as usual residence. All the inhabitants living in the selected dwellings are surveyed.

Following this design, the selection probability of the area h in strata j is

$$p_{hj} = \begin{cases} \frac{A_{hj}}{A_j} \times s_j, & A_{hj} < K_j, \\ 1, & otherwise. \end{cases}$$

where s_j is the number of selected areas in the strata j, A_j is the total number of dwellings in strata j, A_{hj} is the total number of dwellings in the area h and strata j. The selection probability of each dwelling i in area h and strata j is given by

$$p_{ihj} = p_{hj} \times p_{i|hj} = p_{hj} \times \frac{n_{hj}}{A_{hj}} \tag{1.1}$$

where n_{hj} is the number of dwellings sampled in area h and strata j.

Note that in each area h in strata j the selection probability of the dwelling i is constant.

Moreover, since all the individuals in each selected dwelling are surveyed, their selection probabilities are equal to the respective dwelling, $p_{kihj} = p_{ihj}$ for each individual k in dwelling i.

1.2.2 Direct estimation method

The official estimates of the unemployment figures are calculated using a direct method, based on the Horvitz-Thompson estimator (Horvitz and Thompson, 1952). A direct estimator for a specific domain uses only the information of the sampling units in that domain, while an indirect estimator 'borrows strength' through the use of information from the sampling units outside of that particular domain. This information can be entered into the model through appropriated dependence structures, defining the relation between the external information and the domain of interest. Such information can be quite useful for small domains, since their size is not sufficient to produce reliable direct estimates.

Before we proceed with the description of the Horvitz-Thompson estimator (HT), we should first introduce the notation we will use throughout this section. Consider that a finite population with N individuals is divided in m NUTS III regions, $R = R_1, ..., R_j, ..., R_m$. Let N_j be the number of individuals in region j, Y be the interest variable for unemployment estimation, and Y_{jk} the value of Y for the indivual k ($k = 1, ..., N_j$) in the region j ($j = 1, ..., m$), as follows:

$$Y_{jk} = \begin{cases} 1 & \text{if the individual k in region j is unemployed} \\ 0 & \text{otherwise (employed or inactive)} \end{cases} \tag{1.2}$$

Thus, the total of Y, denoted by $Y_{..}$, represents the total unemployed in the population and is given by:

$$Y_{..} = \sum_j \sum_{k \in R_j} Y_{jk}. \tag{1.3}$$

The total unemployed in the region j, $Y_{j.}$, is given by

$$Y_{j.} = \sum_{k \in R_j} Y_{jk}. \tag{1.4}$$

Let s be the sample of the LFS with size n, extracted from the population R through the sampling scheme described in the previous section. Let s_j be a subset of R_j, representing a sample of the NUTS III j.

The HT estimator for the total unemployed is given by

$$\hat{Y}_{..}^{HT} = \sum_j \sum_{k \in s_j} w_{jk} y_{jk},$$

where $w_{jk} = 1/p_{kihj}$ is the inverse selection probability of the individual k in dweling i in area h and strata j, following the sampling scheme in the previous section. w_{jk} is also called 'individual weight'.

Thus, the HT estimator for the total unemployed in the region j is given by

$$\hat{Y}_{j.}^{HT} = \sum_{k \in s_j} w_{jk} y_{jk} \tag{1.5}$$

Let us see that the estimator $\hat{Y}_{j.}^{HT}$ is unbiased, ie, $E[\hat{Y}_{j.}^{HT}] = Y_{j.}$. Note that the estimator can be written as

$$\hat{Y}_{j.}^{HT} = \sum_{k \in R_j} I_{jk} w_{jk} y_{jk},$$

where

$$I_{jk} = \begin{cases} 1 & \text{if } k \in s_j \\ 0 & \text{otherwise} \end{cases}$$

Thus,

$$E[\hat{Y}_{j.}^{HT}] = \sum_{k \in R_j} E[I_{jk} w_{jk} Y_{jk}] = \sum_{k \in R_j} Y_{jk} = Y_{j.}$$

The calculation of the official estimates of the labour market is based on a modified HT estimator, where the weight w_{jk} is based in three factors: an initial weight based on the sampling scheme ($1/p_{kihj}$), a correction for the non-responses, and an adjustment for the known population totals (calibration).

The corrected weight for the non-responses is given by

$$w_{jk}^c = w_{jk} \frac{\hat{N}_j}{\sum_{k=1}^{r_j} w_{jk}} \tag{1.6}$$

where $\hat{N}_j$ is the estimate of population in region R_j where the individual k lives and r_j the number of respondents in region R_j where the individual k lives.

After that correction, an adjustment is made using known totals of population. This adjustment is based on the general regression estimator (GREG), proposed by Deville and Sarndal (1992). The GREG estimator also guarantees the coherence between sampling estimates and known totals of the auxiliary variables as well.

Let us assume that we observe (y_{jk}, x_{jk}) for each individual in the sample s_j, where x is a variable in which the total population $X_j = \sum_{k \in R_j} X_{jk}$ is known. After the correction for the non-responses, the GREG estimator for $Y_{j.}$ is given by

$$\hat{Y}_{j.}^{ds} = \sum_{k \in s_j} w_{jk}^* y_{jk}, \tag{1.7}$$

where w_{jk}^* ($k \in s_j$) are close weights to w_{jk}^c, and are calculated through a distance function subject to the following adjustment condition

$$\sum_{k \in s_j} w_{jk}^* x_{jk} = X_j. \tag{1.8}$$

The idea is to find the closest weight to w_{jk}^c such that the HT estimator for X_j, using the sampling values x_{jk}, coincides with X_j. The optimization problem can be solved through the Lagrange multipliers method, where it is intended to minimize

$$Q(w_1^*, ..., w_n^*, \lambda_j) = \sum_{k \in s_j} D(w_k^*, w_{jk}) - \lambda_j (\sum_{k \in s_j} w_{jk}^* x_{jk} - X_j) \tag{1.9}$$

where D represents the distance function. The NSI uses the logit calibration distance function:

$$D(x) = \left((x - L) log \frac{x - L}{1 - L} + (U - x) log \frac{U - L}{(1 - L)(U - 1)} \right) \frac{1}{A}, \; L < x < U \tag{1.10}$$

where U and L are the upper and lower limits of the calibration weights specified by the user and

$$A = \frac{U - L}{(1 - L)(U - 1)} \tag{1.11}$$

This method garantees positive weights and limits the ratio between the calibration weights and the corrected weights (w_{jk}^*/w_{jk}^c).

Shortcomings

The direct estimator is unbiased. However, the coefficients of variation of this estimator depend strongly on the sample size. In some NUTS III regions, such as Beira Baixa and Terras de Trás-os-Montes, where the sample size is small, the coefficients of variation are quite high. Moreover, for the areas without observations, it is not possible to provide estimates since the direct estimator only uses specific information of the sampling units in the domain of interest. In the LFS, some municipalities do not have any observations at all. Thus, using a direct estimator is not possible to provide estimates for all municipalities.

1.3 Review of small area estimation

In most cases, the sample surveys are used to provide estimates not only for a population of interest, but also for a variety of domains. For example, in this work, domain can be a geographic area or a socio-demographic group. As highlighted before, sometimes the sample size for these domains is not sufficient enough to provide

'direct estimates' (which are based only on the domain-specific area) with adequate precision. Such domains are called 'small areas', and 'small area estimation' (SAE) is the field that deals with this problem. The focus of this field is to study alternative methods, mainly 'indirect estimators' that borrow strenght by using values of the variable of interest from related areas or time periods. Rao and Molina (2015) give a good introduction to this problem and a review of the developed methods to date. The Horvitz-Thompson (HT) estimator and the generalized regression (GREG) estimator, described in section 1.2.2, are the most common direct estimators.

The majority of the SAE methods proposed in the literature are 'small area models'. These models use random effects specific to areas accounting for the variation between those that are not explained by the auxiliary variables. The small area models are classified by Rao and Molina (2015) into two types: area level models that relate the direct estimators to area-specific covariates; and unit level models that relate the unit values of a study variable to unit-specific covariates. Three common estimators can be derived from the small area models: the empirical best linear unbiased prediction (EBLUP), the empirical Bayes (EB) and the hierarchical Bayes (HB). The EBLUP estimator is derived by minimizing the model MSE in the class of linear model unbiased estimators of the quantity of interest. The EB estimator is the conditional expectation of the quantity given the data and the model parameters. The HB estimator is the posterior mean of the estimand, with respect to the posterior distribution of the quantity of interest, given the available data.

One of the most important traditional methods in SAE is the Fay-Herriot (FH) model, proposed by Fay and Herriot (1979). This method is an areal level model that uses the direct estimators as data, instead of the observed values in the sample. By doing this, it can provide directly estimates for the population. Other wise, using the observed data in the sample, the estimates obtained for totals of a given variable of interest, would be adjusted values for the sample and not for the population. In this case, it would be required to use some factor to extrapolate that values from the sample to the population. However, we think that a critical analysis must be done about the assumptions in this model. One of them is the normality assumption for the direct estimators. In many real applications that assumption is not adequate. Moreover, this model considers that the parameter of interest is given by the difference of the direct estimators and an error term with known variance. Thus, this method may add very little to the direct estimates. A technical description of the FH methods in a frequentist approach and in a Bayesian approach will be given below.

1.3.1 Fay-Herriot methods - frequentist approach

The Fay-Herriot model has a standard state space representation with two components: a sampling model for the direct estimates (observation equation, in which the observations are substituted by the direct estimates) and a linking model for the state of the process. The sampling model assumes that a direct estimator $\hat{\theta}_i$ of the parameter of interest θ is available and

$$\hat{\theta}_i = \theta_i + e_i, \quad i = 1, ..., m \tag{1.12}$$

where $e_i's$ are the sampling errors associated with the direct estimator $\hat{\theta}_i$. It is assumed that $e_i's$ are independent normal random variables with mean $E(e_i|\theta_i) = 0$ and variance $Var(e_i|\theta_i) = \sigma_i^2$, which is assumed to be known.

The linking model assumes that the state θ is related to area specific auxiliary data x_i as follows

$$\theta_i = x_i'\beta + v_i, \quad i = 1, ..., m \tag{1.13}$$

where β is the vector of regression coefficients, and the $v_i's$ are random effects specific to areas assumed to be iid with $E(v_i) = 0$ and $Var(v_i) = \sigma_v^2$.

Combining these two models, leads to the FH model

$$\hat{\theta}_i = x_i'\beta + v_i + e_i, \quad i = 1, ..., m \tag{1.14}$$

Note that this model envolves design-induced errors e_i as well as model errors v_i. Usually it is assumed that the sampling variances σ_i^2 are known. However, in practice, it is common to estimate this quantity using a smoothed estimator through the generalized variance function approach (Wolter, 2007). Alternatively, the sampling variance can be modelled directly. See You and Chapmann (2006) for an example of this alternative. You and Zhou (2011) consider both the smoothing and modelling approaches for the sampling variances, applying these models to the analysis of healthy survey data. The BLUP estimator of the small area parameter θ_i, assuming that σ_v^2 is known, is

$$\tilde{\theta}_i = \gamma_i\hat{\theta}_i + (1 - \gamma)x_i'\tilde{\beta}_{WLS} \tag{1.15}$$

where $\gamma_i = \sigma_v^2/(\sigma_v^2 + \sigma_i^2)$ and $\tilde{\beta}_{WLS}$ is the weighted least squared (WLS) estimator of β.

The EBLUP of the small area parameter θ_i based on FH model is obtained by replacing σ_v^2 by the REML estimator in the BLUP estimator.

1.3.2 Fay-Herriot methods - Bayesian approach

Alternatively, Rao and Molina (2015) apply the HB approach to the FH model, assuming a prior distribution on the model parameters. The model proposed is given by

1. Data|Parameters

$$\hat{\theta}_i|\theta_i, \beta, \sigma_v^2 \sim N(\theta_i, \sigma_i^2) \tag{1.16}$$

2. Parameters|Hyperparameters

$$\theta_i|\beta, \sigma_v^2 \sim N(x_i'\beta, \sigma_v^2) \tag{1.17}$$

3. Hyperparameters

$$f(\beta) \propto 1 \tag{1.18}$$
$$f(\sigma_v^2) \sim IG(a_0, b_0) \tag{1.19}$$
$$\tag{1.20}$$

where IG represents the inverse Gamma distribution and a_0, b_0 are chosen to be small, reflecting vague knownledge on σ_v^2. Note that σ_i^2 are supposed to be known.

In recent years, spatial models taking into account spatial dependence among the areas have been proposed in the SAE literature. The most common spatial model in this context is an extension of the FH model. Pratesi and Salvati (2008), Singh *et al.* (2005), and Marhuenda *et al.* (2013) apply spatial FH models using the EBLUP approach (classical approach). You *et al.* (2011), and Jedrzejcczak and Kubacki (2017) apply spatial FH models using the HB approach.

As You *et al.* (2011) explain, a simple spatial extension of the FH model would be to add a spatial random effect u_i in the linking model as follows:

$$\theta_i = x_i'\beta + v_i + u_i \tag{1.21}$$

where $u_i's$ follow the known intrinsic conditional autoregressive model (iCAR), proposed by Besag *et al.* (1991), given as

$$u_i|u_{-i} \sim N\left(\frac{\sum_{j\neq i} w_{ij}u_j}{\sum_{j\neq i} w_{ij}}, \frac{\sigma_u^2}{\sum_{j\neq i} w_{ij}}\right) \tag{1.22}$$

where u_{-i} indicates all the elements in u except the ith, and w_{ij} is 1 if areas i and j are neighbours and 0 otherwise. However, that extension of the FH model has a potential identifiability problem. Only the sum of the random effects $v_i + u_i$ is identified by the data. Thus, here we will consider the parametrization used by You *et al.* (2011), where they consider the model $\theta_i = x_i'\beta + b_i$. They place the following conditional autoregressive model on b:

$$b \sim MVN(0, \sigma_b^2(\lambda R + (1-\lambda)I)^{-1}) \tag{1.23}$$

where σ_b^2 is a spatial disperson parameter, λ is a spatial autocorrelation parameter, I is an identity matrix, and R is the neighbourhood matrix. R has ith diagonal element equal to the number of neighbours of area i, and the other elements equal to -1 if the corresponding areas are neighbours and 0 otherwise. Thus, the HB approach of the model proposed in You *et al.* (2011) is given by

1. Data|Parameters
$$\hat{\theta}|\theta, \beta, \sigma_v^2 \sim MVN(\theta, E) \tag{1.24}$$
where E is a diagonal matrix with the known ith diagonal element σ_i^2.

2. Parameters|Hyperparameters
$$\theta|\beta, \sigma_v^2 \sim MVN(X\beta, \sigma_v^2(\lambda R + (1-\lambda)I)^{-1}) \tag{1.25}$$

3. Hyperparameters
$$f(\beta) \propto 1 \tag{1.26}$$
$$f(\lambda) \sim Uniform(0,1) \tag{1.27}$$
$$f(\sigma_v^2) \sim IG(a_0, b_0) \tag{1.28}$$

where IG represents the inverse Gamma distribution and a_0, b_0 are chosen to be small, reflecting vague knownledge on σ_v^2.

This model reduces to the FH when the spatial autocorrelation parameter $\lambda = 0$. In this case it is assumed independence on the area specific random effects v_i. When $\lambda = 1$, the CAR model 1.23 becames the intrinsic autoregressive model 1.22.

Here, we chose this model with $\lambda = 1$, from the standard SAE methods to do a comparison with the models we will propose on the estimation of unemployment in small areas. The results of the comparative study are presented in the chapter 5.

We also found some SAE literature on an unemployment context. Datta *et al.* (1999) and You *et al.* (2003) propose an extension of FH models to handle time-series and cross-sectional data. Molina *et al.* (2007) and Lopez-Vizcaino *et al.* (2013) proposed area level multinomial mixed models to provide estimates of the three status of the labour market in small areas. Lopez-Vizcaino *et al.* (2015) extend these models, including correlated time random effects.

1.4 Bayesian methods

1.4.1 Basic notions: prior, posterior and predictive distributions

The use of Bayesian methods in statistical analysis has increased over recent years and its modelling can be seen as an extention of the classical paradigm that assigns the model parameters a distribution known as 'prior distribution', by assuming that the parameters themselves are random variables. This prior distribution is then combined with the traditional likelihood to obtain the posterior distribution of the parameter of interest on which the statistical inference is based. The Bayesian philosophy, as Paulino *et al.* (2003) explain, is that the unknown is uncertain and the whole uncertain shall be quantified in terms of probability.

Let $p(y|\theta)$ be the distributional model for the observed data $y = (y_1, ..., y_n)$ given a vector of random parameters $\theta = (\theta_1, ..., \theta_k)$. Notice that $p(.)$ indicates the probability distribution or the density function, according to whether y is discrete or continuous. The data are a random sample from the population. Thus the variability of y depends only on the sampling selection, as Blangiardo *et al.* (2015) explain. The parameter θ is modelled through a prior probability distribution $p(\theta)$. This distribution is defined before we observe y, reflecting the prior belief on θ.

Given the likelihood and prior distribution, Bayes theorem states

$$p(\theta|y) = \frac{p(y|\theta) \times p(\theta)}{p(y)} \tag{1.29}$$

This distribution is called posterior distribution $p(\theta|y)$. The denominator $p(y)$ is the marginal distribution of y, also known as the prior predictive distribution. When θ is a continuous variable $p(y)$ is given by

$$p(y) = \int p(y|\theta)p(\theta)d\theta \tag{1.30}$$

If θ is a discrete variable, it comes

$$p(y) = \sum p(y|\theta)p(\theta) \tag{1.31}$$

Since $p(y)$ does not depend on θ, the posterior distribution $p(\theta|y)$ given in 1.29 can be alternatively reported as

$$p(\theta|y) \propto p(y|\theta) \times p(\theta) \tag{1.32}$$

One of the advantages of using a Bayesian approach is the access to the posterior distribution for the parameter of interest. Indeed, the Bayesian inference is based on that. When $k = 1$, a natural point estimate of θ would be some measure of centrality. The three most common choices are

the posterior mean

$$\hat{\theta} = E(\theta|y) = \int \theta p(\theta|y) d\theta \tag{1.33}$$

the posterior median

$$\hat{\theta} : \int_{-\infty}^{\hat{\theta}} p(\theta|y) d\theta = 0.5 \tag{1.34}$$

and the posterior mode

$$\hat{\theta} : p(\hat{\theta}|y) = sup_{\theta \in \Theta} \quad p(\theta|y) \tag{1.35}$$

The posterior distribution allows us to make any desired probability statments about θ, namely credible intervals. A credible interval can easily be obtained by

$$P(a < \theta < b|y) = \int_a^b p(\theta|y) d\theta = 1 - \alpha \tag{1.36}$$

where $1 - \alpha$ is the desired credible level. Generalization when $k > 1$, although straightforward, is not trivial.

As we can see, the credible intervals have a direct interpretation in terms of probability, contrary to what happens with the frequentist approach for confidence intervals.

In addition to allowing inference regarding an unknown parameter of interest, the Bayesian approach also allows us to make predictions about non-observed variables. When Y_{n+1} is a future observation from the same model as the one which generated $y = (y_1, ..., y_n)$, the posterior predictive distribution is, by definition, given as

$$p(y_{n+1}|y) = E_{\theta|y}[Y_{n+1}] = \int p(y_{n+1}|\theta, y) p(\theta|y) d\theta. \tag{1.37}$$

If Y_{n+1} is stochastically independent of y, then, it simplifies to

$$p(y_{n+1}|y) = \int p(y_{n+1}|\theta) p(\theta|y) d\theta. \tag{1.38}$$

1.4.2 Prior choice

Blangiardo *et al.* (2015) highlight two important aspects which need to be taken into account in the choice of prior distribution: the type of distribution, and the hyperparameters. Usually the chosen type of prior distribution is based on the nature of the parameters of interest. For instance, if the parameter of interest is a proportion, the typical choice is a Beta, because it varies between 0 and 1.

When the posterior distribution belongs to the same family as the prior distribution, that prior is described as being conjugated to the likelihood. This property, called conjugacy, is very convenient. In this case, the functional posterior distribution is known as well as its hyperparameters. Thus, it is easy to derive summary statistics or other quantities of interest.

Conjugate priors exist only for a restricted family of distributions. In most practical applications seldom we can make use of this property. Thus, numerical methods are usually required to perform inference. We will discuss some methods to do this in the next section.

After the specification of the form of the prior distribution, its parameters must be defined according to the informative/noninformative prior knowledge.

When there is a lack of information on the parameters, it is usually recommended to use noninformative priors to allow the data to speak for themselves. There are several ways to construct noninformative priors. One of the most well known noninformative priors is the one proposed by Jeffrey (1946). That prior, which is invariant to transformations, is, for $k = 1$, based on Fisher information:

$$p(\theta) \propto I(\theta)^{1/2} \tag{1.39}$$

where

$$I(\theta) = E\left[\left(\frac{\partial lnp(y|\theta)}{\partial \theta}\right)^2 \middle| \theta\right] \tag{1.40}$$

When $k > 1$, the prior is given by

$$p(\theta) \propto |I(\theta)|^{1/2} \tag{1.41}$$

where $I(\theta)$ is the matrix of Fisher information with each element given by

$$I_{ij}(\theta) = E\left[\frac{\partial lnp(y|\theta)}{\partial \theta_i}\frac{\partial lnp(y|\theta)}{\partial \theta_j}\middle| \theta\right] \tag{1.42}$$

In most cases, it leads to improper prior distributions. This constitutes a problem when the posterior is also improper, which in complex problems may occur, being difficult to check.

Alternatively, the user can adopt difuse or vague priors in the sense that the prior distribution is basically flat on the region of the likelihood where the parameter has more weight. In most cases this is achieved using priors with high variability. For instance, a $Normal(0, 10^6)$ can be used as the prior for a regression parameter, as Blangiardo *et al.* (2015) suggest. A discussion on how to formulate prior distributions can be found, e.g., in Paulino *et al.* (2003).

If, on the other hand, prior information is available, it must be incorporated into the prior distribution. See O'Hagan *et al.* (2006) to understand how the elicitation of experts is done and how it can be incorporated into the models.

The most common choice of priors is nonsubjective as firstly there may be no information of experts and secondly, because for complex hierarchical models the elicitation process is not straightforward. Thus, Simpson *et al.* (2016) suggest a flexible framework called *Penalised Complexity* or 'PC' priors to help the user in specifying informative priors. These priors use only weak information, are straightforward enough to be used by general users, and have a clear meaning and interpretation. We will describe their construction processs in the next section.

1.4.3 Markov chain Monte Carlo methods

As we saw in the previous section, numerical methods based on simulation or approximations are needed when it is not possible to manipulate the posterior distribution in an analytical way.

We will describe some computational methods for Bayesian inference, mainly Markov chain Monte Carlo (MCMC) and integrated nested Laplace approximations (INLA) in the next section.

Before we begin however, it is important to first introduce the Bayesian inference using Monte Carlo methods.

Consider the problem of approximating the posterior mean of a real function of θ, defined as

$$E[g(\theta)|y] = \int g(\theta)p(\theta|y)d\theta \tag{1.43}$$

Given a random sample $\theta_1, ..., \theta_n$ from the posterior density $p(\theta|y)$, the Monte Carlo method approximates the integral in 1.43 by the empirical average:

$$\hat{E}[g(\theta)|y] = \frac{1}{n}\sum_{i=1}^{n} g(\theta_i) \tag{1.44}$$

Moreover, note that the predictive density of a future observation, $p(y_{n+1}|y)$, can be write as $E_{\theta|y}[p(y_{n+1}|\theta, y)]$ (Turkman et al, 2015). Thus, a Monte Carlo approximation for this quantity is given by

$$\hat{p}(y_{n+1}|y) = \frac{1}{n}\sum_{i=1}^{n} p(y_{n+1}|\theta_i, y) \tag{1.45}$$

In both approximations, for the posterior mean and the predictive density, it is necessary to sample directly from the posterior distribution. However, some posteriors have a nonstandard density function to sample from, and others have an unknown form. A possible solution is to draw a sample through a Markov chain whose stationary distribution is the posterior density. The MCMC methods arise from this idea in combination with Monte Carlo methods to compute posterior summaries of interest.

The two most popular MCMC algorithms are the Gibbs sampler and the Metropolis-Hastings algorithm.

We follow Benerjee *et al.* (2015) to briefly describe these methods.

Gibbs sampler

The Gibbs sampler requires that samples can be generated from each of the full conditional distributions $\{p(\theta_i|\theta_{j\neq i}, y), i = 1, ..., k\}$, where $\theta = (\theta_1, ..., \theta_k)'$. The idea of this method is that, under soft conditions, the collection of full conditional distributions uniquely determine the joint posterior distribution $p(\theta|y)$.

To sample from the joint posterior $p(\theta|y)$ the Gibbs sampler draws values iteratively from all conditional distributions. Given an arbitrary set of starting values $\{\theta_2^{(0)}, ..., \theta_k^{(0)}\}$, the Gibbs algorithm is

For $(t \in 1 : T)$, repeat:

1. Sample $\theta_1^{(t)}$ from $p(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, ..., \theta_k^{(t-1)}, y)$
2. Sample $\theta_2^{(t)}$ from $p(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, ..., \theta_k^{(t-1)}, y)$
3. ...
4. Sample $\theta_k^{(t)}$ from $p(\theta_k|\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_{k-1}^{(t)}, y)$

Under general conditions, the realization $(\theta_1^{(t)}, ..., \theta_k^{(t)})$ converges in distribution to the posterior distribution $p(\theta|y)$. Therefore, for t sufficiently large, $\theta^{(t)}, t = t_0 + 1, ..., T$ is a sample from the true posterior, from which any posterior quantities may be estimated.

However, it is important to note that the full conditionals may not have familiar forms, and consequently it may be not possible to sample directly from such distributions. In these cases, alternative methods such as the Metropolis-Hastings (MH) algorithm are usually recommended.

Metropolis-Hastings algorithm

The MH algorithm requires only a function proportional to the distribution to be sampled. Given a starting value $\theta^{(0)}$, the MH algorithm is

For $(t \in 1 : T)$, repeat

1. Sample θ^* from a proposal distribution $q(\theta^*|\theta^{(t-1)})$
2. Compute the ratio

$$r = \frac{p(\theta^*|y)}{p(\theta^{(t-1)}|y)} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{t-1})p(\theta^{t-1})} \tag{1.46}$$

3. Let

$$\theta^{(t)} = \begin{cases} \theta^*, & \text{with probability } min(r, 1), \\ \theta^{(t-1)}, & \text{with probability } 1 - min(r, 1). \end{cases}$$

Under general conditions, a draw $\theta^{(t)}$ converges in distribution to the true posterior density $p(\theta|y)$.

In order to explore the MCMC output and check the convergence, it is strongly recommended to use convergence diagnostic tools.

The implementation of the MCMC methods can be easily performed using, for instance, the software OpenBUGS (Thomas *et al.*, 2006). This software is an open source variant of the WinBUGS (Lunn *et al.*, 2000), a successor to BUGS (Bayesian inference using Gibbs sampling). Alternatively, these methods can be implemented using JAGS (Plummer *et al.*, 2003), Stan (Stan Development Team, 2014) or BayesX (Adler *et al.*, 2013). The first two ones have the advantage of allowing the user to define their own functions and distributions, and the third one is designed for a class of regression models, not as flexible as the previous ones.

For big data, or complex models, the MCMC methods can be extremely slow. Some alternative methods to scale up the MH algorithm have been proposed in the literature of machine learning and computational statistics. Bardenet *et al.* (2017) give a review of these methods and propose a subsampling-based approach relying on a control variate method. One of the alternatives typically used for tall data, is the Divide-and-conquer approach, where the idea is to divide the data into batches, run MH algorithm on each batch separately, and then combine the results. Another alternative is the Pseudo-marginal MH. This method is a variant of MH, which relies on unbiased estimators of an unnormalized version of the target. Bardenet *et al.* (2017) propose an alternative original subsampling approach that, under strong ergodicity assumptions on the original MH sampler, samples from a controlled approximation of the posterior distribution of interest.

However, the methods proposed by Bardenet *et al.* (2017) only perform well in scenarios where the Bernstein-von Mises approximation (van der Vaart, 2000, Chapter 10.2) of the target posterior distribution is excellent.

An alternative method to the MCMC methods, which works in a large range of models and reduces the computational costs, is the INLA, described below.

1.4.4 The integrated nested Laplace approximations

Recently, Rue *et al.* (2009) proposed an alternative method to the MCMC methods, based on the integrated nested Laplace approximations (INLA). This is an analytical approach that allows the user to approximate the posterior distribution of the parameters of interest. Turkman *et al.* (2015) highlight two advantages of this method in comparison with the MCMC methods: the first one is the computational time, and the second is the unified way in which the models are specified. The last point permits one to deal with the inferential process in a more automatized way, independently of the kind of model.

The INLA algorithm was especially designed for latent Gaussian models (LGM). A LGM class can be specified by a hierarchical structure in three levels:

1. Data|Parameters

$$Y|\theta, \psi \sim f(y|\theta, \psi) = \prod_{i=1}^{n} f(y_i|\theta, \psi) \tag{1.47}$$

2. Parameters|Hyperparameters

$$\theta|\psi \sim N(0, Q^{-1}(\psi)) \tag{1.48}$$

In this level it is assumed that the parameter vector θ is modelled by a Gaussian field with Markovian structure (Gaussian Markov Random Field, GMRF), i.e, its density function is given by

$$p(\theta|\psi) = (2\pi)^{-n/2}|Q(\psi)|^{1/2} \exp\left(-\frac{1}{2}\theta^T Q(\psi)\theta\right) \tag{1.49}$$

3. Hyperparameters

$$\psi \sim p(\psi) \tag{1.50}$$

An alternative form to specify the LGM models is through the class of structured additive regression. In this formulation, the dependent variable Y belongs to the exponential family, where the mean μ_i is linked to a predictor η_i with additive structure, through the link function given by $g(\mu_i) = \eta_i$. The general form of the predictor is

$$\eta_i = \beta_0 + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \sum_{k=1}^{n_f} w_{ki} f^{(k)}(u_{ki}) + \epsilon_i \tag{1.51}$$

where β_0 is the intercept, $\beta = (\beta_1, ..., \beta_{n_\beta})$ is the vector with the linear effects of the covariates z. Functions $(f^{(1)}, ..., f^{(n_f)})$ of covariates u can represent non linear effects of continuous covariates, seasonal effects, or structured random effects. w_{ki} can be eventual known weights for each observation. The term ϵ_i can accomodate non structured random effects.

An LGM model is obtained if it is assumed that a GMRF as prior distribution for $\theta = (\beta_0, \beta_j, f^{(k)}(u_{ki}), \eta_i)$.

Notice that the class of LGM is very flexible, accomodating a large range of models.

The objectives of INLA are to obtain analytical approximations for the marginal posterior distributions for the parameters and the hyperparameters of the model. These marginals are given by

$$p(\theta_i|y) = \int p(\theta_i, \psi|y)d\psi = \int p(\psi|y)p(\theta_i|\psi, y)d\psi \tag{1.52}$$

and

$$p(\psi_k|y) = \int p(\psi|y)d\psi_{-k} \tag{1.53}$$

where ψ_{-k} represents the vector ψ without the component ψ_k.

Both the marginals depend on $p(\psi|y)$, which is given by

$$p(\psi|y) = \frac{p(\theta, \psi|y)}{p(\theta|\psi, y)} \tag{1.54}$$

where $p(\theta, \psi|y)$ can be defined as

$$p(\theta, \psi|y) \propto p(\theta|\psi)p(\psi)\prod_i f(y_i|\theta_i, \psi) \tag{1.55}$$

$$\propto p(\psi)|Q(\psi)|^{n/2}\exp\left(-\frac{1}{2}\theta^T Q(\psi)\theta + \sum_i log(f(y_i|\theta_i, \psi))\right) \tag{1.56}$$

Thus, it results

$$p(\psi|y) \propto \frac{p(\theta|\psi)p(\psi)p(y|\theta, \psi)}{p(\theta|\psi, y)} \tag{1.57}$$

$$\approx \left.\frac{p(\theta|\psi)p(\psi)p(y|\theta, \psi)}{\tilde{p}(\theta|\psi, y)}\right|_{\theta=\theta^*(\psi)} \tag{1.58}$$

where $\tilde{p}(\theta|\psi, y)$ is the Gaussian approximation of $p(\theta|\psi, y)$, given by the Laplace method, and $\theta^*(\psi)$ is the mode for a given ψ.

The approximation for $p(\theta_i|\psi, y)$ can be done using one of three approaches:

1. Approximate $p(\theta_i|\psi, y)$ directly as the marginals from $\tilde{p}(\theta|\psi, y)$, using a Normal distribution where the Cholesky decomposition is used for the precision matrix. However, usually that approximation is not very good.

2. Use Laplace approximation to obtain

$$p(\theta_i|\psi, y) = \frac{p(\theta_i, \theta_{-i}|\psi, y)}{p(\theta_{-i}|\theta_i, \psi, y)} \tag{1.59}$$

$$\propto \frac{p(\psi)p(\theta|\psi)f(y|\theta, \psi)}{p(\theta_{-i}|\theta_i, \psi, y)} \tag{1.60}$$

$$\approx \left.\frac{p(\psi)p(\theta|\psi)f(y|\theta, \psi)}{\tilde{p}(\theta_{-i}|\theta_i, \psi, y)}\right|_{\theta_{-i}=\theta^*_{-i}(\theta_i,\psi)} \tag{1.61}$$

   where $\tilde{p}(\theta_{-i}|\theta_i, \psi, y)$ is the Laplace Gaussian approximation to $p(\theta_{-i}|\theta_i, \psi, y)$ and $\theta^*_{-i}(\theta_i, \psi)$ is its mode. This approach works well, but the computational time is expensive.

3. Use the *simplified Laplace approximation*. This approach is based on a Taylor's series expansion of 1.61. Usually, the computational time is short and the approximation is reasonable.

Details about the INLA proceeds for these approximations are given in Blangiardo *et al.* (2015).

1.4.5 Model adequacy and comparison

Checking the model adequacy is very important after any modelling process. In Bayesian modelling, the most commonly used methods are based on predictive distribution. The idea of these methods is to divide the sample in two groups, where one is used to fit the model and the another is used to perform criticism. It can be done using cross-validation or a posterior predictive check.

Cross-validation

The cross-validation used to evaluate the goodness of the model is based on two quantities:

1. the conditional predictive ordinate (CPO) given by $CPO_i = p(y_i|y_{-i})$
2. the probability integral transform (PIT) given by $PIT_i = p(Y_i^* \leq y_i|y_{-i})$

Unusually large or small values of PIT indicate possible outliers. Moreover, a histogram of the PIT value, that is very different from the uniform distribution, indicates that the model is questionable.

The predictive quality of the models can also be analysed using a cross-validated logarithmic score given by the symmetric of the mean of the logarithm of CPO values (Martino and Rue, 2010). High CPO values indicate a better quality of prediction of the respective model.

Posterior predictive check

In the posterior predictive check methods, all the observations are used for model fit and checking. They are based on two quantities:

1. the posterior predictive distribution: $p(y_i^*|y) = \int p(y_i^*|\theta_i)p(\theta_i|y)d\theta_i$
2. the posterior predictive p-value: $p(Y_i^* \leq y_i|y)$

Unusually small values of $p(y_i^*|y)$ indicate observations that can be classified as outliers. If this happens for many observations, the model is not adequate for the data.

If the values of $p(Y_i^* \leq y_i|y)$ are near to 0 or 1, the model is not adequate to fit the data.

Methods based on the deviance

For a comparison between different models, methods based on the deviance are tipically used. Here, we will describe the deviance information criterion (DIC) and the Watanabe-Akaike information criterion (WAIC).

DIC, proposed by Spiegehalter *et al* (2002), is the most commonly used measure of model fit. It is based on a balance between the fit of the model to the data and the corresponding complexity of the model. The fit of the model is measured by the posterior expectation of the deviance

$$D(\theta) = -2\log(p(y|\theta)) + 2\log(h(y)) \tag{1.62}$$

where $p(y|\theta)$ is the likelihood function and $h(y)$ is a standardizing function of the data alone. Usually, it is assumed that $h(y) = 1$ when the models being compared have the same sampling distribution.

The complexity of the model is measured by the effective number of parameters

$$p_D = E_{\theta|y}(D(\theta)) - D(E_{\theta|y}) = \bar{D} - D(\bar{\theta}) \tag{1.63}$$

Thus, the DIC is given by

$$DIC = \bar{D} + p_D \tag{1.64}$$

The model with the smallest value of DIC is the one with a better balance between the model adjustment and complexity. However, this criterion can present some problems, which arise in part from not being fully Bayesian.

A typical alternative is the WAIC, proposed by Watanabe (2010), which is fully Bayesian in that it uses the entire posterior distribution. The WAIC is given by

$$WAIC = -2\sum_{i=1}^{n} \log(E_{\theta|y}(p(y_i|\theta))) + 2p_W \tag{1.65}$$

where $p_W = \sum_{i=1}^{n} Var_{\theta|y}(\log(p(y_i|\theta)))$. This criterion can be considered as an improvement on the DIC for Bayesian models (Gelman *et al.*, 2014). Again, smaller values of WAIC indicate a better model.

1.5 A brief summary of areal and point referenced data methods and models

Spatial statistics has generated significant attention in many areas, such as epidemiology, engineering and environmental health among many others. Data with a spatial nature not only provide information about the attributes of interest, but also about the geographical location of these attributes. That information must be incorporated into the model for the data. As Tobler (1970) states in a famous sentence known as the first law of geography, 'Everything is related to everything else, but near things are more related than distant things'. The spatial dependence can be incorporated through structured spatial random effects. The specification of the models depends on whether or not the data are areas or points.

Cressie (1991) suggests a very useful tool to understand and perform spatial data analysis. He describes spatial data as resulting from observations on the stochastic process

$$\{Y(s) : s \in D\} \tag{1.66}$$

where D is possibly a random set of $\Re^d$. Cressie (1991), Banerjee *et al.* (2004) and Blangiardo *et al.* (2015) make a distinction between three types of spatial data:

1. *Areal data*, where D is a fixed subset partitioned into a finite number of areas with well defined boundaries;

2. *Spatial point patterns*, where D is itself random. $Y(s)$ can simply equal 1 for all $s \in D$ (indicating the occurrence of the event), or possibly give some additional information (marked point pattern process);

3. *Geostatistical data*, where D is fixed and $Y(s)$ is a random vector at a location $s \in \Re^d$, where s varies continuously over D.

The understanding of this distinction is crucial for our work. We had the opportunity to analyse each of the three types of unemployment spatial data, as they became available. At the beginning, only the unemployment data in areas of mainland Portugal was available, so we used areal data instead. From the 4th quarter of 2014, all the sampled dwellings in the LFS are georeferenced, thus we analysed spatial point patterns (the locations of the population units are random). After that, the georreferencing of all dwellings in the country became available to our study. Since all locations are now known, we are able to use geostatistical data.

The following chapters address these three data analyses. Since the spatial detail of the data and the temporal period are distinctive in the three approaches, we will describe the data used in each chapter.

Before we describe the methodologies we are proposing, we will first provide a brief introduction of the three types of models.

1.5.1 Areal data models

The modelling of spatial data must incorporate the spatial dependence. In the areal data models, that dependence can be considered through random effects based on the spatial structure of the neighbourhood. However, sometimes the inclusion of random effects in a model with non-normal likelihood can induce some complexity in the model. The most effective way of handling this is through hierachical bayesian modelling. Using that modelling approach, the structure of a generalized linear model may be specified in three levels:

1. data|link function
2. link function|parameters, latent terms
3. parameters|hyperparameters

One of the most well developed subjects in this field is disease mapping, which we will consider next.

Disease mapping is a field of strong epidemiological interest that focuses on the main areas of Bayesian hierarchical modelling and its application to the analysis of disease. Lawson (2009) presents a good overview on the topic, and in this context, the data are totals of disease or deaths in each area and time. Let y_{jt} be the number of observed cases and E_{jt} be the number of expected cases in area j and time t. The standard spatio-temporal model in disease mapping is given by

$$y_{jt} \sim Poisson(E_{jt}\rho_{jt}) \tag{1.67}$$

with

$$\log(\rho_{jt}) = \alpha_0 + u_j + v_j + \gamma_t + \phi_t \tag{1.68}$$

where α_0 is the intercept, v_j is the area-specific effect modelled as exchangeable, u_j is another area-specific effect modelled as spatially structured, ϕ_t is the time-specific effect modelled as exchangeable, and γ_t is another time-specific effect modelled as spatially structured.

The usual assumptions are

1. u is modelled as an *intrinsic conditional autoregressive* (iCAR) process, proposed by Besag *et al.* (1991), with the following specification

$$u_i|u_{-i} \sim Normal(\frac{1}{N_i}\sum_{j=1}^{n} a_{ij}u_j, s_i^2) \tag{1.69}$$

   where u_{-i} indicates all the elements in u except the ith, N_i is the set of neighbours of area i, a_{ij} is 1 if areas i and j are neighbours and 0 otherwise, and $s_i^2 = \sigma_u^2$ is the variance for area i.

2. γ is modelled as a first order *random walk* (AR (1)) defined as

$$\gamma_t|\gamma_{t-1} \sim Normal(\gamma_{t-1}, \sigma^2) \tag{1.70}$$

Knorr-Held (2000) expand this model to allow for an interaction between time and space, using the following linear predictor:

$$\log(\rho_{jt}) = \alpha_0 + u_j + v_j + \gamma_t + \phi_t + \delta_{jt} \tag{1.71}$$

where the specification of δ_{jt} depends on the effects that we assume to interact. See Knorr-Held (2000) and Blangiardo *et al.* (2015) for details.

1.5.2 Spatial point processes

Diggle (2003) defines a spatial point pattern as a set of locations, irregularly distributed within a designated region and presumed to have been generated by some form of stochastic mechanism. The main objective of point process statistics is to understand the spatial structure of these patterns. Unlike classical statistics, point process statistics are confronted with various types of correlation in the patterns. As Illian *et al.* (2008) state, the distances between the points are correlated, as well as the number of points in adjacent regions. The eventual marks attached to the points may also be correlated. The nature of a spatial pattern can be described using appropriate statistical methods.

In general, a pattern can be classified as regular, random or aggregated. When the occurrence of an event at a particular location makes it more likely that other events will occur nearby, we say that the pattern is aggregated. In contrast, when the occurrence of an event makes it less likely we say that the pattern is regular. When the locations are independent from each other, we say that the pattern is random.

The most basic model for point process modelling is the Poisson process. A homogeneous Poisson process has two properties: first that the density of points is constant, and second that the locations are independent from each other. This is called complete spatial randomness. However, usually these assumptions are not valid in practical problems.

Diggle (2003) describes some generalizations of the homogeneous Poisson process, such as:

1. Poisson cluster processes - were introduced by Neyman and Scott (1958) to model clustered patterns. This kind of pattern is very common in real data applications. The definition of these processes is based on three postulates: parent events form a Poisson process; each parent produces a random number of offspring, realized independently and identically for each parent; and that the positions of the offspring relative to their parents are independently and identically distributed.

2. Inhomogeneous Poisson processes - are obtained by replacing the constant intensity λ of the Poisson process with a spatially varying intensity function $\lambda(x)$. This class has the following properties: $N(A)$ has a Poisson distribution with mean $\int_A \lambda(x)dx$, and given $N(A) = n$, the n events in A form an independent random sample from the distribution on A with pdf proportional to $\lambda(x)$.

3. Cox processes - are a particular case of inhomogeneous Poisson processes, where the intensity is stochastic. The Cox processes have two properties: $\{\Lambda(x) : x \in \Re^2\}$ is a non-negative-valued stochastic process, and conditional on $\{\Lambda(x) = \lambda(x) : x \in \Re^2\}$ the events form an inhomogeneous Poisson process with intensity function $\lambda(x)$. A more flexible and tractable version of Cox processes are the log-Gaussian Cox processes (LGCP), where it is assumed $\Lambda(x) = \log(Z(x))$, where $Z(x)$ is a Gaussian field.

4. Simple inhibition processes - are commonly used for modelling regular patterns, by the imposition of a minimum permissible distance between any two events.

5. Markov point processes - are a more flexible class for regular patterns than the simple inhibition processes, since they permit competitive interactions between the objects. For example, due to the competitive interaction, two individuals can survive in close proximity to each other.

Here, we will focus on the log Gaussian Cox processes since the georeferenced unemployment data present a clustered pattern. The specification of the Gaussian field required in these models leads to computational problems in the inference, due to the complexity of the covariance structure (known as 'big n problem', Blangiardo et al, 2015). To solve this problem, we will use the *Stochastic Partial Differential Equations* (SPDE) methodology, introduced by Lindgren *et al.* (2011), which permits us to approximate a Gaussian field (continuous process) by a Gaussian Markov random field (discrete process). By doing this, we can adopt the INLA approach.

Before the description of the SPDE methodology and the LGCP models, we introduce the concepts of Gaussian random fields and Gaussian Markov random fields.

A random field $\{Z(s), s \in D \subset \Re^d\}$ is a stochastic process where s are the locations and $Z(s)$ is a random variable observed at locations $s \in D$. If the random vector $Z(s)$ has multivariate Gaussian distribution with mean μ and covariance matrix Σ, it is called a Gaussian random field.

A Gaussian Markov random field arises in a discrete domain, such as a regular grid or a collection of spatial locations. The random vector $Z = (Z_1, ..., Z_n)^T$ with mean μ and precision matrix $Q > 0$ is a Gaussian Markov random field if, and only if, $Z \sim N(\mu, \Sigma = Q^{-1})$ and $Q_{ij} \neq 0 \iff i, j$ are neighbours. One example of this process is the iCAR model defined in 1.69.

Stochastic Partial Differential Equations approach

The Stochastic Partial Differential Equations (SPDE) approach was proposed by Lindren *et al.* (2011) to represent a gaussian field (GF) using a Gaussian Markov random field (GMRF).

This methodology uses a computational mesh for representing the latent Gaussian random field. The following finite element representation is assumed

$$Z(s) \approx \sum_{j=1}^{N} w_j \psi_j(s) \tag{1.72}$$

where N is the number of the mesh nodes, $w = (w_1, w_2, ..., w_N)^T$ is a multivariate Gaussian random vector (representing a Gaussian Markov random field, GMRF) and $\{\psi_j\}_{j=1}^{N}$ are the selected base functions defined for each mesh node: ψ_j is 1 at mesh node j and 0 in all other mesh nodes. w is chosen so that the distribution of $W(s)$ approximates the distribution of the solution to the SPDE given by

$$(k - \Delta)^{\alpha/2}(\tau Z(s)) = \Lambda(s) \tag{1.73}$$

where $s \in \Re^d$, Δ is the Laplacian, α controls the smoothness, $k > 0$ is the scale parameter, τ controls the variance, and $\Lambda(s)$ is a Gaussian spatial white noise process.

Lindgren *et al* (2011) showed that the resulting distribution for the weights, which is the solution of the SPDE, is $w \sim N(0, Q(\tau, k)^{-1})$ where the precision matrix $Q(\tau, k)$ is a polynomial in the parameters τ and k. Working directly with the SPDE parameters k and τ can be difficult because they both affect the variance of the field (Yuan *et al* (2017)). So, we will consider the standard deviation σ and the spatial range ρ which are given, respectively, by

$$\sigma = \sqrt{\frac{1}{4\pi k^2 \tau^2}} \tag{1.74}$$

and

$$\rho = \frac{\sqrt{8}}{k} \tag{1.75}$$

Log-Gaussian Cox process model

A log-Gaussian Cox process (LGCP) $N = \{s, s \in D\}$ is a Cox process with intensity function given by

$$\lambda(s) = \exp(Z(s)) \tag{1.76}$$

where $\{Z(s), s \in D\}$ is a Gaussian random field.

A possible re-parametrization of the intensity of the LGCP is given by

$$\lambda(s) = \exp\{\beta + Z(s)\} \tag{1.77}$$

Moreover, an alternative version, described in Pereira *et al.* (2014) is given by

$$\lambda(s) = \exp\{\beta_0 + \beta^T X(s) + Z(s)\} \tag{1.78}$$

where β_0 and β are the regression coefficients and $X(s)$ are the spatial covariates.

Conditional on a realization of $Z(s)$, a log-Gaussian Cox process is an inhomogeneous Poisson process. It follows therefore that the likelihood for an LGCP is of the form

$$\log(p(\theta|\mathbf{x})) = |\Omega| - \int_\Omega \lambda_1(s|\mathbf{x},\theta)ds + \sum_{s_i \in S} \lambda_1(s_i|\mathbf{x},\theta), \tag{1.79}$$

where S is the set of observed locations and $\lambda_1(s)$ is defined in (1.76).

The integral in the likelihood is intractable due to the stochastic nature of $\lambda_1(s)$. To solve this problem, we could use the traditional methods to fit a log-Cox process, which consist of dividing the study regions into cells, forming a lattice, and then counting the number of points into each one. These counts are modelled using the Poisson likelihood. See for example Illian *et al* (2010). However, Simpson *et al* (2016) consider that this approach can be very inefficient, especially when the intensity of the process is high, the window of observation is too large or when the pattern is rare. Instead, they propose the use of an SPDE approach, introduced by Lindgren *et al* (2011), to transform a Gaussian field (GF) to a Gaussian Markov random field (GMRF), as we explained in the previous section. Notice that this methodology uses a computational mesh only for representing the latent Gaussian random field and not for modelling the counts.

Using that approximation, it follows that the integral in (1.79) can be written as

$$\int_\Omega \lambda_1(s)ds = \int_\Omega \exp(Z(s))ds \approx \int_\Omega \exp\left(\sum_{j=1}^{N} w_j \psi_j(s)\right) ds \tag{1.80}$$

This integral can be approximated using standard numerical integration schemes. Simpson *et al* (2016) suggest using the following quadrature rule

$$\int_\Omega f(s)ds \approx \sum_{i=1}^{N+n} \beta_i f(s_i) \tag{1.81}$$

where $\{s_i\}_{i=1}^{N+n}$ are the locations of mesh nodes and observations, and $\{\beta_i\}_{i=1}^{N+n}$ are the quadrature weights.

Unlike the traditional methods for inference in LGCP models, this methodology uses each location to model the point pattern, without aggregation. The LGCP model belongs to the group of latent Gaussian models, and consequently, the inference can be performed within the INLA platform.

Marked point processes

Following the definition given in Illian *et al.* (2008), marked point processes are models for random point patterns where marks that describe properties of the objects are attached to the points. Here, we will denote a marked point process by $M = [x_n; m(x_n)]$, where $m(x_n)$ is the mark of the point x_n.

Thus, in a marked point pattern, the objects are characterized not only by their positions but also by marks, i.e, additional data on each individual object. The marks can be quantitative or qualitative.

Illian *et al.* (2008) describe three types of marked point process models:

1. Independently marked point process - this model assumes independent marks.
2. Random field model - this model assumes that correlated marks are obtained from the random field model, which is independent of the point process.
3. Intensity-weighted marks - a model in which the point density and marks are correlated. Both are modelled by a common random field.

We will consider the three approaches in chapter 3 in order to select the most adequate model for our data.

Numerical challenges and solutions

Some numerical challenges arise when we are implementing a spatial point process model using the SPDE approach. One such complication is choosing the most suitable mesh to be used in the GMRF representation. The mesh is a carefully constructed collection of triangles that must cover the entire spatial domain of interest. The R-INLA package contains some functions that can be used to construct a mesh. It requires some information as input, such as details about the spatial domain (this could be the locations of observations or the domain extent) and the largest triangle edge length. It is also possible to extend the triangulation outside of the domain to avoid any boundary effect where we have a variance much larger than within the domain itself. Furthermore, it is important to do some sensitivity analysis using different meshes. The more triangles, the more precise are the representation of the GMRF, but, more computational time is needed in the modelling process. Thus, the choice of the mesh is a balance between the accuracy and the computational costs. Krainski *et al.* (2017) give an R-INLA tutorial on SPDE models with a good explanation of the construction of a mesh. The R-code used in this thesis can be found in the appendix.

Another challenge that can arise is choosing of the priors for the SPDE parameters. We followed Simpson *et al* (2017) and Fuglstad *et al* (2017) to construct a joint penalising complexity (PC) prior density for the spatial range, ρ, and the marginal standard deviation, σ, which is given by

$$p(\rho, \sigma) = RS\rho^{-2}e^{-R\rho^{-1}-S\sigma} \tag{1.82}$$

where R and S are hyperparameters determined by $R = -\log(\alpha_1)\rho_0$ and $S = -\log(\alpha_2)/\sigma_0$.

The practical approach for this in INLA is to require the user to indirectly specify these hyperparameters through $P(\rho < \rho_0) = \alpha_1$ and $P(\sigma < \sigma_0) = \alpha_2$.

Furthermore, the availability of auxiliary information at the locations of the observations and mesh nodes required in the modelling of an LGCP was also challenging. We had some auxiliary information for the sampling units and intended to extrapolate that for the mesh nodes. To do this, we used a Kernel smoothing method. The R-code used is also in the appendix. We used the function 'smooth.ppp' from the package 'spatstat' developed by Baddeley et al (2016). The idea behind this method, also known as the Nadaraya-Watson smoother (Nadaraya, 1964, 1989; Watson, 1964) is the following: if the observed values are $y(s_1), ..., y(s_n)$ at locations $s_1, ..., s_n$ respectively, then the smoothed value at a location u is given by

$$g(u) = \frac{\sum k(u - s_i)y(s_i)}{\sum k(u - s_i)} \tag{1.83}$$

where k is a probability density. A common choice for this is the isotropic Gaussian probability density and in this case it is called a Gaussian kernel. The standard deviation of the kernel is the smoothing bandwidth, that can be specified in the function 'smooth.ppp'. A larger bandwidth gives more smoothing. As Baddeley et al (2016) explain, for very large values of smoothing bandwidth, the result will be approximately constant and equal to the average mark value in the entire dataset. For small values of bandwidth, the result becomes closer to the nearest data point. The choice of the bandwidth involves a balance between the bias and variance: as bandwidth increases, the bias increases and variance decreases. In the last chapter, we make a sensitivity analysis using different values for the bandwidth parameter.

Some other numerical problems arose during the modelling process. One of which was related to the coordinates system used, as an LGCP model can be sensitive to this. In this case, a projection from the latitude and longitude to the utm system in km is recommended. In addition, due to the complexity of an LGCP model, some convergence problems can arise when sampling from an approximated posterior of the fitted model. In addition, the criterion WAIC may become unstable. To avoid these numerical problems, we suggest the use of the following instruction 'control.inla = list(int.strategy = "eb"))' in the call of the 'inla' function. We also suggest the command 'inla.rerun' after the inla call, to make the model more stable.

1.5.3 Model based geostatistics

The main problem in geostatistics is the prediction of a variable of interest over a domain, based on the values observed at a limited number of points.

The most classical approach to spatial prediction in the point-referenced data setting is the *kriging* (Diggle *et al.*, 1998). Let $Y = (Y(s_1), ..., Y(s_n))'$ be the observations of a random field, and let us assume that we intend to predict the variable Y at a site s_0 where it has not been observed. Consider the following linear predictor $Y(s_0) = \sum l_i Y(s_i) + \delta_0$. The idea of kriging is to minimize the mean squared error, $E[(Y(s_0) - (\sum l_i Y(s_i) + \delta_0)^2]$, over δ_0 and l_i.

In the context of Gaussian processes, the kriging is based on the following model

$$Y = \mu + \epsilon, \tag{1.84}$$

where

$$\epsilon \sim N(0, \Sigma) \tag{1.85}$$

and

$$\Sigma = \sigma^2 H(\psi) + \tau^2 I \tag{1.86}$$

where $H_{ij}(\phi) = \rho(\phi, d_{ij})$, $d_{ij} = |s_i - s_j|$, ρ is a valid correlation function on $\Re^r$, and τ^2 is the nugget effect variance.

When covariate values $x = (x(s_1), ..., x(s_n))'$ and $x(s_0)$ are available for incorporation into the analysis, the model takes the form

$$Y = X\beta + \epsilon, \tag{1.87}$$

where

$$\epsilon \sim N(0, \Sigma) \tag{1.88}$$

Banerjee *et al.* (2015) show that the predictor that minimizes the error is the conditional expectation of $Y(s_0)$ given the data.

Let $\theta = (\beta, \sigma^2, \tau^2, \phi)$ be the vector of model parameters. In a Bayesian context, the model above can be written as

$$Y|\theta \sim N(X\beta, \sigma^2 H(\phi) + \tau^2 I) \tag{1.89}$$

$$p(\theta|y) \propto f(y|\theta)p(\theta) \tag{1.90}$$

Alternatively, we can add a vector of spatial random effects W:

$$Y|\theta, W \sim N(X\beta + W, \tau^2 I) \tag{1.91}$$

$$W|\sigma^2, \phi \sim N(0, \sigma^2 H(\phi)) \tag{1.92}$$

The model specification is completed by adding priors for β, τ^2, σ^2 and ϕ.

Extension of krigging to non-normal distribution

For many types of data, the normal distribution is not adequate. By analogy with generalized linear models, here we can use generalized linear spatial process models. Banerjee *et al.* (2015) describe a generic model, where the Gaussian model is replaced by another member of the class of exponential family models.

$$Y|\theta \sim f(Y|\theta) \tag{1.93}$$

where

$$g(\theta) = X^T\beta + W \tag{1.94}$$

where $g(.)$ is an adequate link function, and

$$W \sim N(0, \sigma^2 H(\phi)) \tag{1.95}$$

If the number of observations is large, the modelling approach involving the spatial covariance can be time consuming and unstable. Thus, we use the SPDE approach, such as in the previous section.

Chapter 2

Areal data modelling

2.1 Introduction

As we described in the section of Small Area Estimation, the majority of the SAE methods are areal data models. In a context of unemployment the SAE literature is quite limited. Datta *et al.* (1999) and You *et al.* (2003) proposed a spatio temporal extension of the FH model for unemployment rates. However, the FH model assumes normality for the data, which does not seem to be adequate in this case. Here, generalized linear models would be preferable. Thus, Da-Silva and Migon (2016) suggest a hierarchical dynamic beta model for modelling unemployment rates and proportions. Molina *et al.* (2007) and Lopéz-Vizcaíno *et al.* (2015) suggest a multinomial model for modelling the three status of the labour market (employement, unemployement, inactivity).

Here, we consider three different modelling strategies: the modelling of the total number of unemployed people through the Poisson, Binomial, and Negative Binomial models; modelling the unemployment rate using a Beta model; and the simultaneous modelling of the total of the three categories of the labour market (employment, unemployment and inactivity) using a Multinomial model.

The inference will be made in a bayesian context using the R-INLA package of software R.

2.2 Data

The region under study (mainland Portugal) is partitioned into 28 NUTS III regions (NUTS-2002), indexed by $j = 1, ..., 28$. We did not include the autonomous regions because they coincide with the NUTS II regions for which estimates are already available with acceptable accuracy.

We use the Portuguese Labour Force Survey data from the 1st quarter of 2011 to the 4th quarter of 2013 in order to produce accurate estimates for the labour market indicators in the last quarter. Each quarter is denoted by $t = 1, ..., 12$. We did not use more recent data because there was a change in the sampling design during 2014 and that could affect the temporal analysis.

We are interested in the total unemployed population, and the unemployment rate of the population by NUTS III regions, which is denoted by Y_{jt} and R_{jt}. We

denote the respective sample values by y_{jt} and r_{jt}. The unemployment rate is given by the ratio of active people who are unemployed, as defined by the European regulation of the Labour Force Survey.

The models developed to make estimation in small areas gain special importance with the inclusion of variables of interest, which we call covariates. In this study, some potential covariates were identified and following that a selection using the stepwise method and an analysis of the correlations was conducted. The selected covariates are divided into 5 groups: population structure, economy, labour market, companies and type of economic activity. Some of these covariates are regional and are static in time whereas others are available per quarter and thus are also of dynamic nature. We will make the distinction and classify these sets of covariates into regional, temporal and spatio-temporal covariates. These selected covariates are as follows:

a) Population structure: a.1) Proportion of individuals in the sample of the Labour Force Survey that are female and aged between 24 and 34 years (F_24_34, regional and quarterly); a.2) Proportion of individuals in the sample of the Labour Force Survey that are female and over 49 years (F_49, regional and quarterly);

b) Economy: b.1) Gross domestic product per capita (GDP, quarterly);

c) Labour market: c.1) Proportion of unemployed people registered in the employment centers (IEFP, regional and quarterly);

d) Companies: d.1) Number of enterprises per 100 inhabitants (regional);

e) Type of economic activity: e.1) Proportion of population employed in the primary sector of activity (regional); e.2) Proportion of population employed in the secondary sector of activity (regional).

Figure 2.1 shows the evolution of the unemployment rate observed in the sample from the Portuguese Labour Force Survey from the 1st quarter of 2011 to the 4th quarter of 2013 in each of the 28 NUTS III . The bold represents the average unemployment rate. We can see that for all regions there was a slight increase in the unemployment rate during this period.

The map in Figure 2.2 shows the spatial and temporal distribution of the unemployment rate observed in the sample of Portuguese Labour Force Survey during the period under study. As we can see, this map suggests the existence of spatial and temporal dependence structures in the observed data.

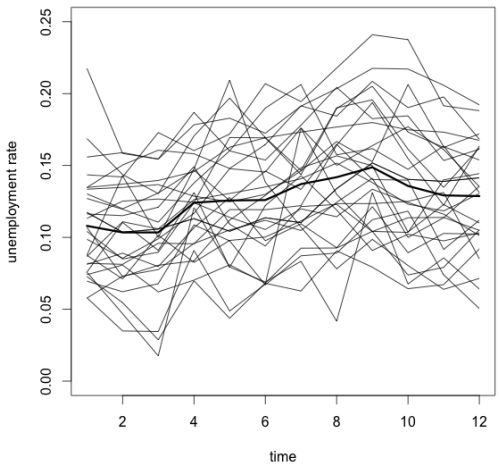


Figure 2.1: Unemployment rate observed in the sample from the Portuguese Labour Force Survey from the 1st quarter of 2011 to the 4th quarter of 2013 in each of the 28 NUTS III

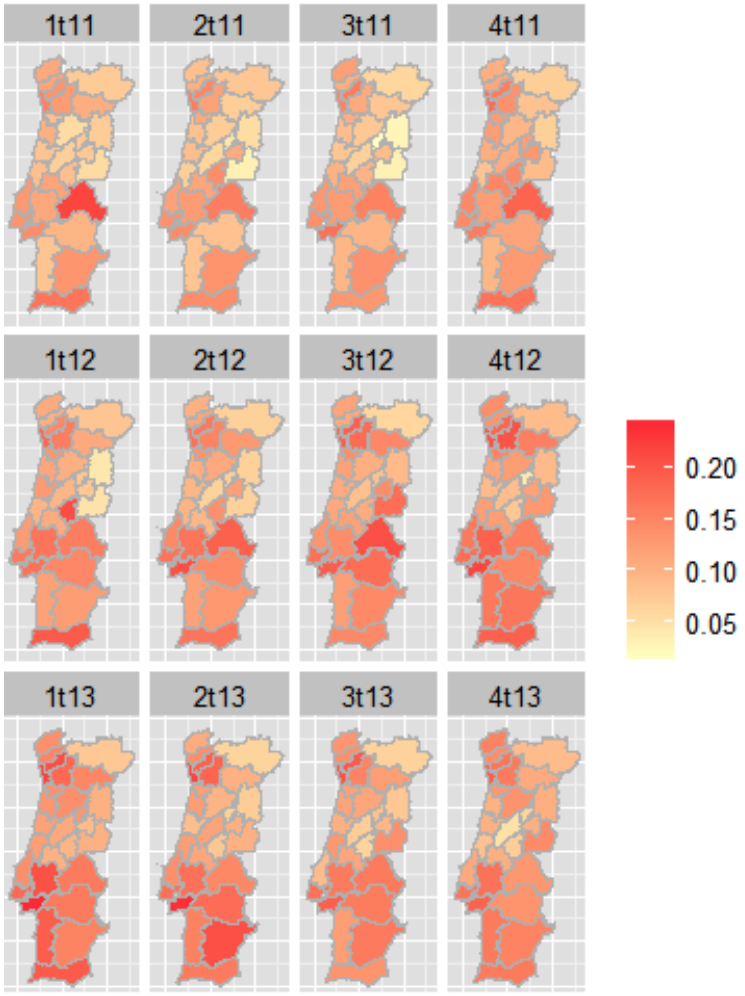


Figure 2.2: Spatial and temporal distribution of the unemployment rate observed in the sample of the Portuguese Labour Force Survey.

2.3 Bayesian models for counts and proportions

In this problem we are interested in estimating the effect of selected variables on the number of unemployed individuals and the unemployment rate, taking into account the temporal and spatial correlations.

One of the most general and useful ways of specifiying this problem is to employ hierarchical generalized linear model set up, in which the data are linked to covariates and spatial-temporal random effects through an appropriately chosen likelihood and a link function which is linear on the covariates and the random effects.

We denote the vector of designated regional covariates by $\mathbf{x_j} = (x_{1j}, x_{2j}, x_{3j})$, the temporal covariates by x_t and the vector of spatio-temporal covariates by $\mathbf{x_{jt}} = (x_{1jt}, x_{2jt}, x_{3jt})$.

While modelling unemployment numbers, we generically assume that

$$y_{jt}|\mu_{jt} \sim \pi(y_{jt}|\mu_{jt}), \quad j = 1, ..., 28, \quad t = 1, ..., 12,$$

where π is a generic probability mass function. We look at this model considering specific probability mass functions, such as Poisson and Binomial, among others. The state parameters μ_{jt} depend on covariates and on structured and unstructured random factors through appropriate link functions.

The unemployment rate is also hierarchically modelled in a similar way. We assume that

$$r_{jt}|\theta_{jt} \sim g(r_{jt}|\theta_{jt}), \quad j = 1, ..., 28, \quad t = 1, ..., 12,$$

where g is a properly chosen probability density function and θ_{jt} are the state parameters.

In the following sections we look at different variations of these hierarchical structures with different link functions.

Let us consider h, the chosen link function which depends on the assumed model for the data. We assume $\eta_{jt} = h(\mu_{jt})$ for the modelling of the total and $\eta_{jt} = h(\theta_{jt})$ for the modelling of the rates. For each model, we consider the following linear predictor

$$\eta_{jt} = offset_{jt} + \alpha_0 + \mathbf{x_j'}\boldsymbol{\alpha} + x_t'\beta + \mathbf{x_{jt}'}\boldsymbol{\gamma} + w_{jt} + \epsilon_{jt}, \quad j = 1, ..., 28 \quad t = 1, ..., 12, \quad (2.1)$$

where $offset_{jt}$ are constants that can be included in the linear predictor during adjustment. The vectors $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$, β and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$ correspond respectively to vectors of the covariates coefficients $\mathbf{x_j}$, x_t and $\mathbf{x_{jt}}$. Components ϵ_{jt} represent unstructured random effects, which assume

$$\epsilon_{jt} \sim N(0, \sigma_\epsilon^2),$$

and the components w_{jt} represent the structured random effects that can be written as $w_{jt} = w_{1j} + w_{2t}$ where $\mathbf{w_1}$ is modelled as a *intrinsic conditional autoregressive* (ICAR) process proposed by Besag *et al.* (1991) and $\mathbf{w_2}$ is modelled as a first order *random walk* (AR (1)). Blangiardo *et al.* (2015) succinctly describe both the ICAR and AR (1) processes.

$$\mathbf{w_1}|\tau_{w_1} \sim ICAR(\tau_{w_1}),$$
$$\mathbf{w_2}|\tau_{w_2} \sim AR(1).$$

We assume the following prior distributions for the regression parameters

$$\begin{aligned}
&\alpha_0 \sim N(0, 10^6),\\
&\alpha_i \sim N(0, 10^6) \quad i = 1, 2, 3,\\
&\beta \sim N(0, 10^6),\\
&\gamma_i \sim N(0, 10^6) \quad i = 1, 2, 3.
\end{aligned}$$

For the hyperparameters we assume

$$\begin{aligned}
&\log \tau_\epsilon \sim \log Gamma(1, 0.0005),\\
&\log \tau_{w_1} \sim \log Gamma(1, 0.0005),\\
&\log \tau_{w_2} \sim \log Gamma(1, 0.0005).
\end{aligned}$$

We assume the following models for the distribution of the observed data: Poisson, Binomial, and Negative Binomial for the total of unemployed, Beta for the unemployment rate and Multinomial for the total of the three status of the labour market (employment, unemployment and inactivity).

2.3.1 Poisson model

This is perhaps the most frequently used model for counting data in small areas, especially in epidemiology. If we consider that μ_{jt} is the mean of the total number of unemployed people, we can assume that

$$y_{jt}|\mu_{jt} \sim Poisson(\mu_{jt}), \quad j = 1, ..., 28, \quad t = 1, ..., 12.$$

Therefore

$$p(y_{jt}|\mu_{jt}) = \mu_{jt}^{y_{jt}} \exp(-\mu_{jt})/y_{jt}!, \quad y_{jt} = 0, 1, 2...$$

In this case, the link function is the logarithmic function ($\log = h$). The NUTS III regions have different sample dimensions, so the variation of the total level of unemployment is affected. To remove this effect, we need to add an $offset$ term, which is given by the number of individuals in the sample in each NUTS III region.

2.3.2 Negative Binomial model

The Negative Binomial model may be used as an alternative to the Poisson model, especially when the sample variance is much higher than the sample mean. When this happens, we say that there is over-dispersion in the data. In this case, we can assume that

$$y_{jt}|\mu_{jt},\phi \sim \textit{Negative Binomial}(\mu_{jt},\phi), \quad j=1,...,28 \quad t=1,...,12.$$

The probability mass function is given by

$$p(y_{jt}|\mu_{jt},\phi)=\frac{\Gamma(y_{jt}+\phi)}{\Gamma(\phi).y_{jt}!}.\frac{\mu_{jt}^{y_{jt}}.\phi^{\phi}}{(\mu_{jt}+\phi)^{y_{jt}+\phi}}, \quad y_{jt}=0,1,2...$$

where $\Gamma(.)$ is the gamma function.

The most convenient way to connect μ_{jt} to the linear predictor is through the $\log\frac{\mu_{jt}}{\mu_{jt}+\phi}$. Also in this case, the term $offset$ described in the Poisson model is considered.

2.3.3 Binomial model

When measuring the total number of unemployed people, we may also consider that there is a finite population in the area j. In this case, we assume that this population is the number of active individuals in the area j, which is denoted by m_{jt}, assuming that it is fixed and known. We can then consider a Binomial model for the total number of unemployed given the observed active population. So, given the population unemployment rate R_{jt},

$$y_{jt}|m_{jt},R_{jt} \sim \textit{Binomial}(m_{jt},R_{jt}), \quad j=1,...,28 \quad t=1,...,12,$$

which means that

$$p(y_{jt}|m_{jt},R_{jt})=\binom{m_{jt}}{y_{jt}}R_{jt}^{y_{jt}}(1-R_{jt})^{m_{jt}-y_{jt}}, \quad y_{jt}=0,1,...,m_{jt},, \quad t=1,...,12.$$

In this case, the most usual link function is the logit function given by $\log(R_{jt}/(1-R_{jt}))$.

We expect that the fit of this model will be close to the fit of the Poisson model in the regions with a big number of active people and a small unemployment rate.

2.3.4 Beta model

The Beta distribution is one of the most commonly used model for rates and proportions. We can assume that the unemployment rate r_{jt} follows a Beta distribution and using the parameterization proposed by Ferrari and Cribari-Neto (2004), we denote by

$$r_{jt}|\mu_{jt},\phi \sim \textit{Beta}(\mu_{jt},\phi), \quad j=1,...,28 \quad t=1,...,12.$$

The probability mass function is given by

$$p(r_{jt}|\mu_{jt},\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu_{jt}\phi)\Gamma((1-\mu_{jt})\phi)}r_{jt}^{\mu_{jt}\phi-1}(1-r_{jt})^{(1-\mu_{jt})\phi-1}, \quad 0<r_{jt}<1,$$

where $0 < \mu_{jt} < 1$ and $\phi > 0$.

In this case, there are several possible choices for the link function, but the most common is the logit function $h(\mu_{jt}) = \log(r_{jt}/(1 - r_{jt}))$.

2.3.5 Multinomial model

The Multinomial logistic regression model is an extension of the Binomial logistic regression model and is used when the variable of interest is multi-category. In this case, it may interest us to model the three categories of the labour market (employment, unemployment and inactivity), giving us the unemployment rate which can be expressed by the ratio between the unemployed and the active people (the sum of the unemployed and employed).

One advantage of the Multinomial model in this problem is the consistency obtained between the three categories of the labour market. The estimated total employment, unemployment and inactivity coincides with the total population. In addition, the same model provides estimates for the rate of employment, unemployment and inactivity.

Assuming that $\mathbf{y_{jt}} = (y_{jt1}, y_{jt2}, y_{jt3})$ is the vector of the total in the three categories of the labour market, the Multinomial model can be written as

$$\mathbf{y_{jt}}|n_{jt}, \mathbf{P_{jt}} \sim Multinomial(n_{jt}, \mathbf{P_{jt}}), \quad j = 1, ..., 28 \quad t = 1, ..., 12,$$

where n_{jt} is the number of individuals in the area j and quarter t, and $\mathbf{P_{jt}} = (P_{jt1}, P_{jt2}, P_{jt3})$ is the vector of proportions of employed, unemployed and inactive, where $P_{jt3} = 1 - (P_{jt1} + P_{jt2})$.
The probability mass function is given by

$$p(y_{jt1}, y_{jt2}, y_{jt3}|n_{jt}, P_{jt}) = \frac{n_{jt}!}{y_{jt1}!y_{jt2}!y_{jt3}!} P_{jt1}^{y_{jt1}} P_{jt2}^{y_{jt2}} P_{jt3}^{y_{jt3}},$$

where

$$y_{jtq} \in \mathbb{N} : \sum_{q}^{3} y_{jtq} = n_{jt}, \ q = 1, 2, 3.$$

The most common link function is the log of P_{jtq}, defined as $\eta_{jtq} = \log(P_{jtq}/P_{jt3})$, $q = 1, 2$.

2.4 Application to the Portuguese LFS data

2.4.1 Results

This section provides the results of applying five models for the estimation of the total number of unemployed people and unemployment rate to the NUTS III regions of Portugal.

The Poisson, Binomial, Negative Binomial and Beta models were implemented using the R package *R-INLA*, while the Multinomial model was implemented based on MCMC methods using the R package *R2OpenBUGS*.

When the Multinomial regression model was combined with the predictor given in 2.1, some convergence problems arose, due to its complexity. For this reason, the effects w_{jt} and ϵ_{jt} were replaced by the unstructured area and time effects u_j and v_t, where it was assumed

$$u_j \sim N(0, \sigma_u^2),$$
$$v_t \sim N(0, \sigma_v^2),$$

with the following prior information

$$\log \tau_u \sim \log Gamma(1, 0.0005),$$
$$\log \tau_v \sim \log Gamma(1, 0.0005).$$

Due to the differences in the model structure and the computational methods used for the Multinomial model, the comparative analysis of results for this model should be done with some extra care.

The posterior mean of the parameters and hyperparameters of each model as well as the standard deviation and the quantile 2.5 % and 97.5 % are presented in Tables 2.1, 2.2, 2.3, 2.4 and 2.5. We can see that the covariates *GDP* and *secondary sector* are not significant for any of the models applied. However, the value obtained for *Deviance Information Criterion* (DIC) increases considerably without the inclusion of these variables, so we decided to include them.

We observe that the IEFP is significant in all of the models applied, as expected. The number of enterprises per 100 000 inhabitants has a negative effect on the increase of unemployment. The population structure has also a significant effect. The proportion of individuals that are female and aged between 24 and 34 years has a positive effect on the increase of unemployment. On the other hand, the proportion of individuals that are female and over 49 years has a negative effect. These tendencies are probably due to young unemployment in the first case and to the fact that the age group +49 includes most of the inactive people, in the second case.

All the considered models give very good fit to the data and their temporal predictions are also satisfactory. Here we report on several model fitting aspects of the Binomial model. Similar results for the other models are given in the Supplementary Material.

Figure 2.3 a) gives the observed and adjusted values from the Binomial model together with their 95% credible intervals, whereas figure 2.3 b) gives the predictions to the 4th quarter of 2013 together with their 95% credible intervals. We see that the adjusted values are very close to the observed ones. The domains are sorted at first by quarter and then by region. This is the reason for the identical behavior in each 28 domains (corresponding to the NUTS III regions). The graphs show a slight increase on the unemployment rate until the 1st quarter of 2013 and then a decrease until the 4th quarter of 2013.

The map of the figure 2.4 allows for a better understanding of the regional difference between the observed and fitted values.

Poisson				
Parameter	Mean	SD	2.5Q	97.5Q
(Intercept)	-2.83	0.01	-2.85	-2.81
Companies	-0.01	0.02	-0.05	0.02
Primary sector	-0.02	0.72	-1.45	1.40
Secondary sector	0.02	0.21	-0.39	0.43
GDP	0.00	0.00	0.00	0.00
IEFP	10.05	0.96	8.17	11.93
F_24_34	4.30	1.34	1.65	6.93
F_49	-1.55	0.57	-2.65	-0.42
τ				
τ_{w_2}	25047.76	20819.39	3297.22	79744.21
τ_{w_1}	25.77	9.52	11.91	48.78
τ_ϵ	22082.79	19692.19	2213.88	73957.91

Table 2.1: Posterior mean, standard deviation and 95% credible intervals for the parameters and hyperparameters of the Poisson model.

Negative Binomial				
Parameter	Mean	SD	2.5Q	97.5Q
(Intercept)	-2.83	0.01	-2.86	-2.81
Companies	-0.01	0.02	-0.05	0.02
Primary sector	0.13	0.73	-1.33	1.57
Secondary sector	-0.04	0.23	-0.48	0.41
GDP	0.00	0.00	0.00	0.00
IEFP	10.20	1.48	7.28	13.09
F_24_34	3.97	2.01	0.02	7.91
F_49	-2.11	0.73	-3.54	-0.67
τ	48.57	5.44	38.69	60.05
τ_{w_2}	22946.14	20085.32	2453.14	75579.11
τ_{w_1}	32.77	14.26	13.08	68.00
τ_ϵ	22641.11	19924.30	2405.53	74957.89

Table 2.2: Posterior mean, standard deviation and 95% credible intervals for the parameters and hyperparameters of the Negative Binomial model.

Binomial				
Parameter	Mean	SD	2.5Q	97.5Q
(Intercept)	-1.97	0.01	-2.00	-1.95
Companies	-0.04	0.02	-0.07	0.00
Primary sector	0.54	1.01	-1.47	2.52
Secondary sector	-0.11	0.28	-0.67	0.45
GDP	0.00	0.00	0.00	0.00
IEFP	12.63	1.11	10.47	14.81
F_24_34	4.38	1.47	1.50	7.26
F_49	-1.11	0.64	-2.37	0.16
τ				
τ_{w_2}	20736.79	19243.68	2070.46	71553.26
τ_{w_1}	11.77	3.90	5.70	20.85
τ_ϵ	19143.06	18555.71	1460.60	68268.97

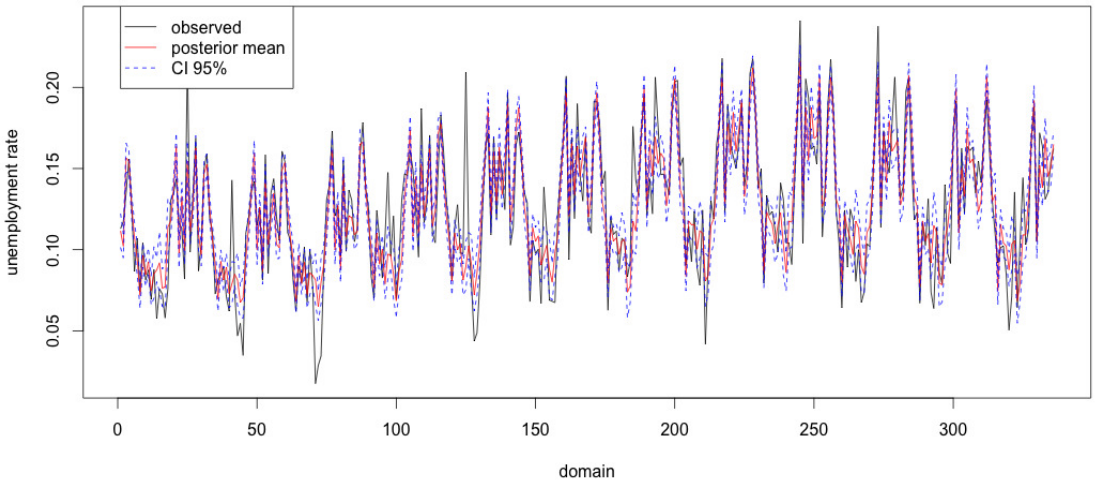
Table 2.3: Posterior mean, standard deviation and 95% credible intervals for the parameters and hyperparameters of the Binomial model.

Beta				
Parameter	Mean	SD	2.5Q	97.5Q
(Intercept)	-1.98	0.01	-2.00	-1.95
Companies	-0.03	0.03	-0.08	0.02
Primary sector	0.69	1.09	-1.50	2.83
Secondary sector	-0.20	0.31	-0.82	0.42
GDP	0.00	0.00	0.00	0.00
IEFP	12.22	1.82	8.64	15.78
F_24_34	0.85	1.84	-2.77	4.47
F_49	-2.37	0.68	-3.70	-1.03
τ	206.61	16.80	174.43	240.37
τ_{w_2}	20012.58	19075.29	1750.73	70491.89
τ_{w_1}	11.33	4.61	5.40	23.02
τ_ϵ	20497.48	19404.83	1715.97	71768.00

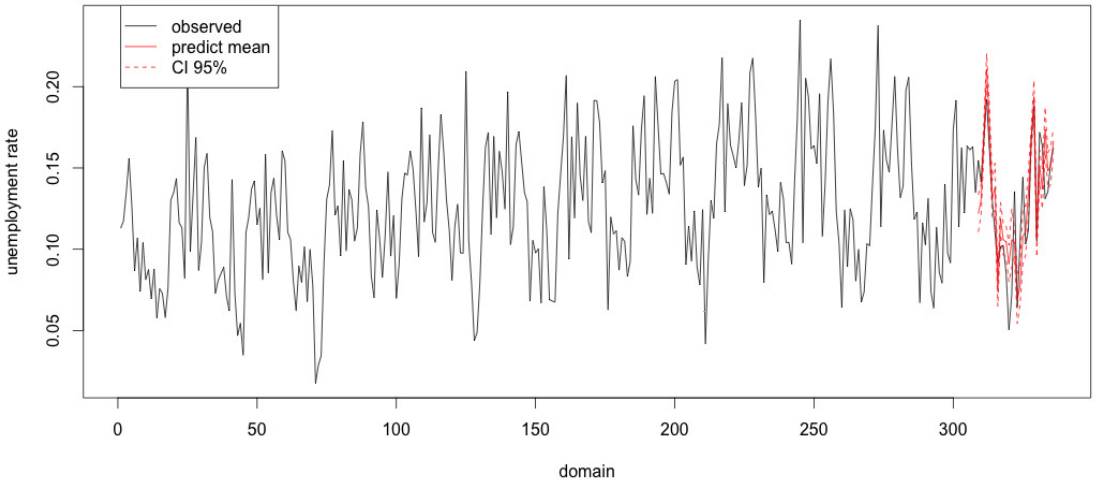
Table 2.4: Posterior mean, standard deviation and 95% credible intervals for the parameters and hyperparameters of the Beta model.

Multinomial				
Parameter	Mean	SD	2.5Q	97.5Q
(Intercept)	-1.74	0.26	-2.16	-1.20
Companies	0.01	0.02	-0.04	0.05
Primary sector	3.99	5.38	-1.04	14.94
Secondary sector	-0.77	0.77	-2.27	0.19
GDP	0.00	0.00	0.00	0.00
IEFP	8.93	1.59	6.02	12.00
F_24_34	4.77	1.44	1.97	7.58
F_49	-2.38	0.64	-3.74	-1.22
τ_v	2206.25	2979.60	3.32	9519.00
τ_u	33.57	24.87	1.77	78.11

Table 2.5: Posterior mean, standard deviation and 95% credible intervals for the parameters and hyperparameters of the Multinomial model.



(a)



(b)

Figure 2.3: a) Observed and adjusted values (mean and 95 % CI) of the unemployment rate for the 336 domains (336 = 28 NUTS III × 12 quarters); b) Observed and predicted values (the posterior mean and 95 % CI) of the unemployment rate. The prediction is made for the 4th quarter of 2013 which is highlighted in red.

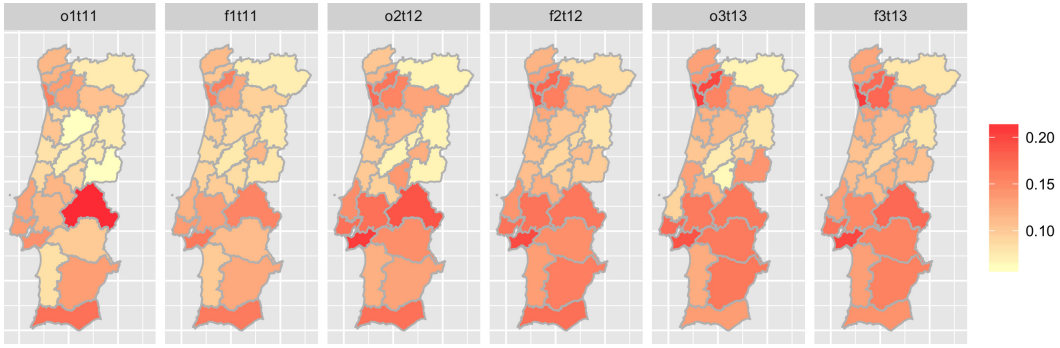


Figure 2.4: Maps of observed and fitted values of the unemployment rate for the 1st quarter of 2011, 2nd quarter of 2012 and 3rd quarter of 2013.

2.4.2 Diagnosis

Some predictive measures can be used for an informal diagnostic, such as *Conditional Predictive ordinates* (CPO) and Probability Integral Transforms (PIT; Gelman et al, 2004). Measure CPO_i is defined as $\pi(y_i|y_{-i})$ where y_{-i} is the vector y without observation y_i, while the measures PIT_i are obtained by $Prob(y_i^{new} \leq y_i|y_{-i})$. Unusually large or small values of this measure indicate possible outliers. Moreover, a histogram of the PIT value which is very different from the uniform distribution indicates that the model is questionable.

The implementation of these measures in an MCMC approach is very heavy and requires a high computational time. For this reason, we present only results for the models implemented with the INLA.

Figure 2.5 shows the graphs of the PIT values versus domain ($28 \times 12 = 336$) and the histogram of the PIT values for Poisson model. The histogram for the PIT values based on the Poisson and Binomial models presents a fairly uniform behavior, but this is not the case with the Negative Binomial and Beta distributions. This suggests that these last two models may not be suitable for data in analysis.

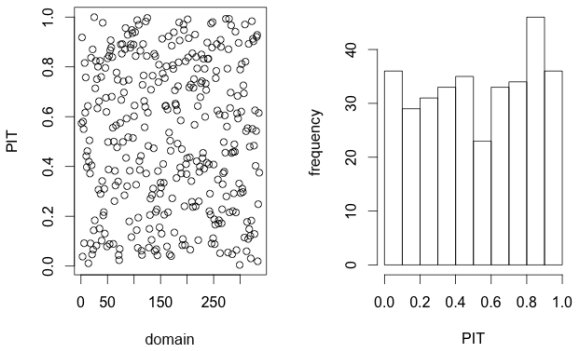


Figure 2.5: Graphs of the PIT values versus domain ($28 \times 12 = 336$) and the histogram of the PIT values.

The predictive quality of the models can be performed using a cross-validated logarithmic score given by the symmetric of the mean of the logarithm of CPO values (Martino and Rue, 2010). High CPO values indicate a better quality of prediction of the respective model. The logarithmic of the CPO values are given in table 2.6. Accordingly, the Beta model has the least predictive quality.

The diagnosis of the Multinomial model was based on graphical visualization and on *Potential Scale Reduction Factor* (Brooks and Gelman, 1997). No convergence problems were detected.

Model	log score
Poisson	3.33
Negative Binomial	3.51
Binomial	3.34
Beta	-2.39

Table 2.6: Logarithmic score

2.5 Comparison between the estimates for the proportion of unemployment using the Binomial model and the traditional SAE methods

Let us consider that we are interested in the estimation of the proportion of unemployment by NUTS III. In order to make a comparison between the results obtained using the Binomial model and the direct method, as well as the SAE methods described in the first chapter (FH and FH-CAR), we employ each of them using data from the LFS in the 4th quarter of 2014. We use the same set of covariates in the FH, FH-CAR, Binomial and Binomial CAR models.

Figure 2.7 shows the estimates obtained from each model for NUTS III regions. Grande Lisboa and Grande Porto are the most populated regions, thus we expect that the direct method will perform well here. Notice that the Binomial and Binomial CAR models disagree with the direct method and the traditional SAE methods in these regions. This result may indicate some bias in the estimates obtained by the models. Figure 2.6 shows the direct estimates for the proportion of unemployment and the 95% credible intervals obtained by the Binomial CAR model for the NUTS III regions. Indeed, the estimates obtained by the Binomial model seem to be biased, since the direct estimator is unbiased and its estimates are not inside the intervals for some regions. However the coefficientes of variation (CVs), given by the ratio between the mean to the standard deviation, favour these models. As we can see in figure 2.8, the models with the best performance in terms of variability are the FH and the Binomial models without structured random effects. These models present naturally lower variances but large biases. Notice that in the regions with lower population density, the direct method presents high CVs, as we expected. The FH-CAR model improved the results of the direct method, but the Binomial CAR model presents even better results in terms of variability in the regions with low population density.

The relative performance of the Binomial and Binomial CAR models was as we expected. The estimators considered are naive synthetic estimators, since they are based on the product of the domain sample count of unemployed people prediction and the ratio between the population size and the sample size in that domain. Usually, these kinds of estimators have low variances and large biases. The FH models avoid this problem as they model the direct estimates for the population. Therefore, they take into account the selection probabilities of the dwellings in the sample of the LFS at km^2 level. The generalized linear models proposed here take into account these probabilities at NUTS III level, leading to high bias. In

the following chapters, we will explore some alternative approaches that solve this problem.

Although the FH models produced low biases in this case, the normal distribution is not naturally adequate for modelling proportions.

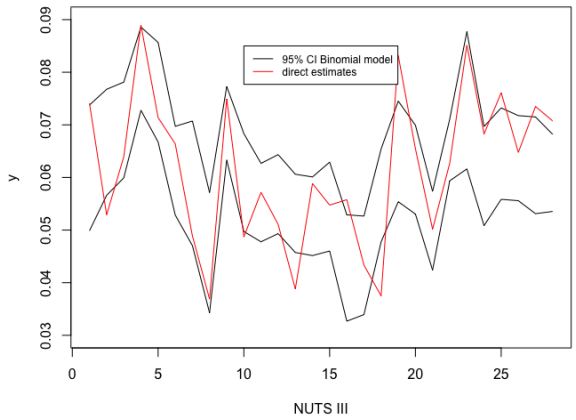


Figure 2.6: Direct estimates and the 95% credible intervals obtained by the Binomial CAR model

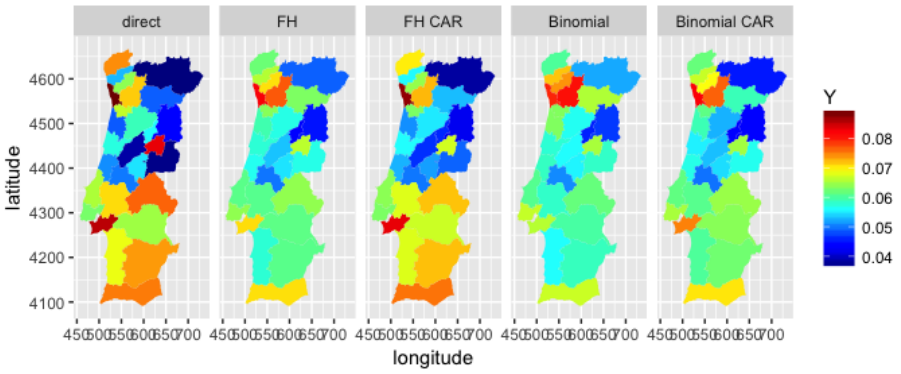


Figure 2.7: Estimates of the proportion of unemployment by NUTS III for the 4th quarter of 2014 using the direct method, the FH model, the FH-CAR model, the Binomial model, and the Binomial CAR model

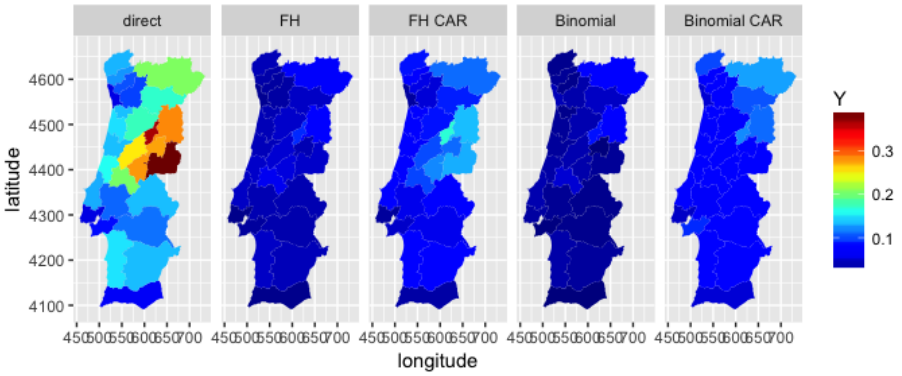


Figure 2.8: Coefficients of variation of the estimates using the direct method, the FH model, the FH-CAR model, the Binomial model, and the Binomial CAR model

2.6 Comparison between the estimates for the total number of unemployed people using the Poisson model and the traditional SAE methods

Let us consider now that we are interested in the estimation of the total number of unemployed people. In order to make a comparison between the estimates obtained using the Poisson model and the direct method, as well as the SAE methods, we employ each of them using data from the LFS in the 4th quarter of 2014. We use the same set of covariates in the FH, FH-CAR, Poisson and Poisson CAR models.

Figure 2.10 shows the estimates for the total number of unemployed people obtained from each model for NUTS III regions. Here we can see that the Poisson CAR model is the closest model to the direct method in the highest population regions, which is a favorable result. Figure 2.9 confirms that the estimates obtained by the Poisson CAR model are approximatly unbiased (the direct estimates are inside the 95% credible intervals in all regions). In terms of variability, the FH CAR model performs better than the Poisson CAR model (see figure 2.11). When the domains of interest have different population sizes, we do not recommend the use of FH models for totals. In this case, a Poisson model can take into account the population size in the modelling process through the offset term. FH models ignore that information, resulting in biased estimates.

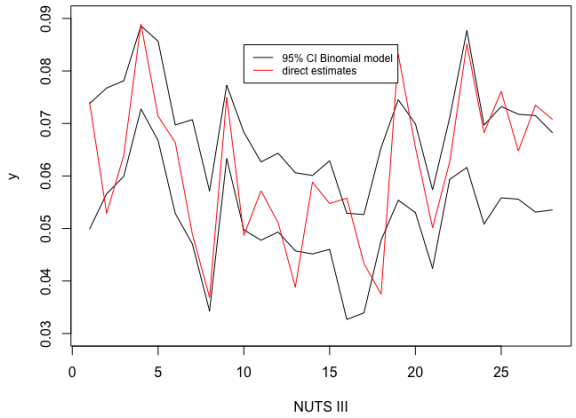


Figure 2.9: Direct estimates and the 95% credible intervals obtained by the Poisson CAR model

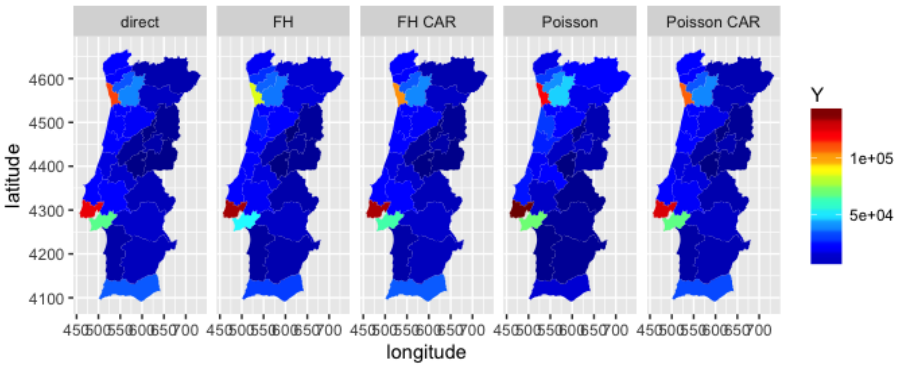


Figure 2.10: Estimates of total unemployed by NUTS III for the 4th quarter of 2014, using the direct method, the FH model, the FH-CAR model, the Poisson model, and the Poisson CAR model

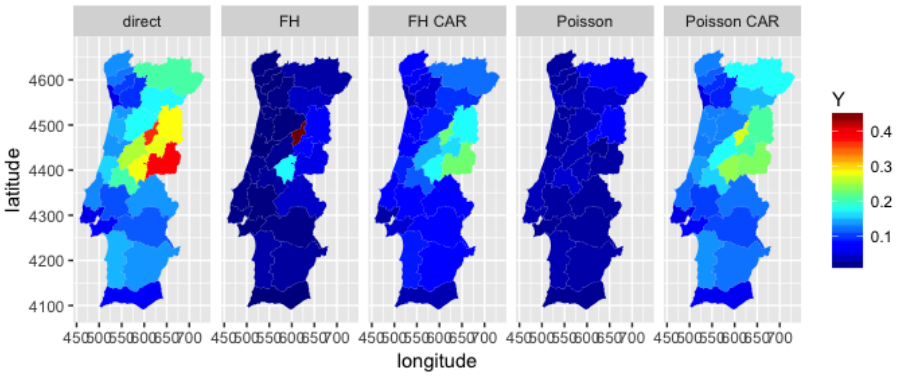


Figure 2.11: Coefficients of variation of the estimates using the direct method, the FH model, the FH-CAR model, the Poisson model, and the Poisson CAR model

Chapter 3

Spatial point processes modelling

3.1 Introduction

From the 4th quarter of 2014 onwards, all the sampling units (dwellings) of the LFS were georeferenced according to the coordinates of the respective buildings. Geo-referencing residential units permits using methods and models for point referenced data, increasing the spatial resolution of inferential methods and providing much more detailed information on the intensity of unemployment across space. This approach allows for the representation of the sample survey as a realization of a spatial point process together with the associated marks, namely the number of unemployed people in each residential unit. Point referencing also permits us to use more precise auxiliary information at residential units, such as the average education level or income of their inhabitants.

Note that a residential building may be composed of one or more dwellings. Therefore, different sampling units may have the same spatial location causing some difficulty in generating modelling strategies. We will address this problem by moving away from dwellings to residential buildings as our sampling units. This strategy will solve the awkward problem of multiplicity of geo-referenced units with the cost of introducing some approximations at residential unit level and some consequential loss of precision.

For modelling the intensity of residential unit locations and their associated marks, we suggest using marked Log Gaussian Cox processes (LGCP) as a model. The LGCP is a class of flexible models widely used in the context of spatial point processes (Moller and Waagepetersen, 2003, Illian *et al*, 2008, Baddeley *et al*, 2016). Typically, in this framework, the log intensity of the point process is assumed to be a (latent) Gaussian random field. In order to facilitate calculations, often marks are assumed to be independent of point patterns so that marks and spatial patterns can be modelled separately. However, for inference on such models, Illian *et al* (2012) proposed a flexible framework using INLA (Rue *et al*, 2009, Martins *et al*, 2013, Rue *et al*, 2017), in which the spatial patterns of points and marks are allowed to be dependent, and assume their independence conditional on common latent spatial Gaussian processes, making these models more flexible. Inference on such models is not straightforward. Due to computational problems that emerge in this framework, Lindgren *et al* (2011) proposed a more computationally tractable approach

based on stochastic partial differential equation (SPDE) models, which permit the transformation of a Gaussian field to a Gaussian markov random field. This is the method that we decided to follow.

Diggle (2003) defines a spatial point pattern as a set of locations, irregularly distributed within a designated region and presumed to have been generated by some form of stochastic mechanism. The main objective of point process statistics is to understand the spatial structure of these patterns. Unlike classical statistics, point process statistics are confronted with various types of correlation in the patterns. As Illian *et al.* (2008) state, the distances between the points are correlated, as well as the number of points in adjacent regions. The eventual marks attached to the points may also be correlated. The nature of a spatial pattern can be described using appropriate statistical methods.

In general, a pattern can be classified as regular, random or aggregated. When the occurrence of an event at a particular location makes it more likely that other events will occur nearby, we say that the pattern is aggregated. In contrast, when the occurrence of an event makes it less likely we say that the pattern is regular. When the locations are independent from each other, we say that the pattern is random.

The most basic model for point process modelling is the Poisson process. A homogeneous Poisson process has two properties: first that the density of points is constant, and second that the locations are independent from each other. This is called complete spatial randomness. Diggle (2003) describes some generalizations of the homogeneous Poisson process, such as the inhomogeneous Poisson process and the Cox processes. The inhomogeneous Poisson processes are obtained by replacing the constant intensity λ of the Poisson process with a spatially varying intensity function $\lambda(x)$. This class has the following properties: N(A) has a Poisson distribution with mean $\int_A \lambda(x)dx$, and given N(A)=n, the n events in A form an independent random sample from the distribution on A with pdf proportional to $\lambda(x)$. In particular, the intensity can be stochastic. In this case, the processes are called Cox processes. The Cox processes have two properties: $\{\Lambda(x) : x \in \Re^2\}$ is a non-negative-valued stochastic process, and conditional on $\{\Lambda(x) = \lambda(x) : x \in \Re^2\}$ the events form an inhomogeneous Poisson process with intensity function $\lambda(x)$. A more flexible and tractable version of Cox processes are the log-Gaussian Cox processes (LGCP), where it is assumed $\Lambda(x) = \log(Z(x))$, where $Z(x)$ is a Gaussian process. We will focus on this process.

3.2 Data

In this study, we analyze the quarterly data of the Labour Force Survey (LFS) regarding to the 4th quarter of 2014. The sample size is about 40,000 observations and each individual can be classified into one of the following three categories: employed, unemployed and inactive. Covariate information about the individuals are available, such as gender, age and education level.

Within a point process modelling scheme, the choice of dwellings as sampling units, creates problems. Since residential units are geo-referenced, multiple sampling units appear with the same spatial location. The sampling design and the consequent

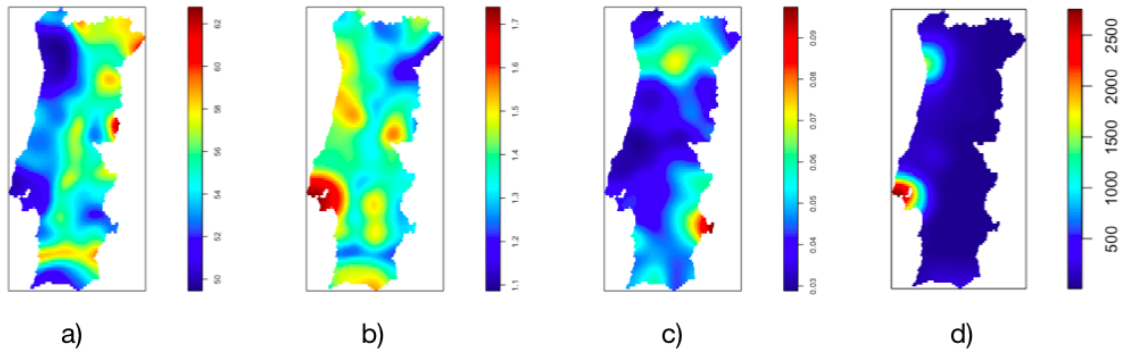


Figure 3.1: Kernel estimates of: a) the mean age per residential building; b) the median of the education level per residential building; c) the proportion of registered unemployed people in centers of employment; d) the population density (inhabitants/km^2).

data are not sufficiently detailed to obtain good information on the multiplicity distribution of dwellings in each residential building, therefore we use residential buildings as design units. Thus we aggregate the number of unemployed observed in dwellings with the same spatial location and we denote by $Y(s_j)$, the number of unemployed in the residential building at spatial position s_j.

3.2.1 Covariates

One of the great benefits of point referenced models for spatial variation of unemployment is that very detailed covariate information can be given at residential building level, like median of education and mean age of the individuals in each residential building. For the locations we have available only one covariate, the population density, which we will use as offset in the model.

The median of the education level in each residential building and the mean age were considered as covariates to model the marks. Although the education level does not constitute a quantitative variable, it was treated as such due to its ordinal meaning (1-primary level, 2-secondary level, 3-higher level). Higher values of this variable in the Grande Lisboa and Península de Setúbal regions can be clearly seen (figure 3.1). It is also interesting to see the spatial distribution of the mean age, with more younger people near the country's coastline than in its interior. The proportion of unemployed people registered in the employment centers depends on the number of unemployed and this information is also used as a covariate. Kernel based estimates of the covariates are given in figure 3.1.

3.3 A spatial point patterns approach

We model the spatial point process $N_1(\cdot)$ of residential units by a log Gaussian Cox process with intensity $\lambda_1(s)$, with

$$\log \lambda_1(s|W(s)) = \alpha_1 + z_1'(s)\theta_1 + W(s), \tag{3.1}$$

We assume that for every s, the mark $Y(s)$ is a Poisson random variable with probability mass function

$$P_{Y(s)|W(s)}(y) \sim \text{Poisson}(\lambda_2(s|W(s))), \tag{3.2}$$

where

$$\log \lambda_2(s|W(s)) = \alpha_2 + z_2'(s)\theta_2 + \alpha_3 W(s). \tag{3.3}$$

Here, $W(s)$ is a latent Gaussian Markov field, $z_1(s)$, $z_2(s)$ are generic auxiliary information which may help in understanding the spatial patterns of points as well as the marks and $\theta = (\alpha_1, \alpha_2, \alpha_3, \theta_1, \theta_2)$ are the model parameters.

We assume that the same latent Gaussian process $W(s)$ is acting both on the point patterns and their marks with different scaling and we assume that conditional on $W(s)$, point mass density and the marks are independent. It is natural to expect that in areas with high density of residential buildings, one would expect higher rate of unemployment and therefore independence of points and marks may not be an reasonable assumption.

With the conditonional independence assumption, the corresponding marked point process $N(s, y)$ has the following structure:

$$N_1(s, y)|\lambda(s, y) \sim \text{Poisson}(\lambda(s, y)), \tag{3.4}$$

with

$$\lambda(s, y|W(s)) = \lambda_1(s|W(s))P_{Y(s)|W(s)}(y) \tag{3.5}$$

where $\lambda_1(s|W(s))$, and $P_{Y(s)|W(s)}(y)$ defined as in (3.1),(3.2) and (3.3).

3.3.1 Target quantities for inference

Our objective is to make inference on the number of unemployed in any given region A, based on the sample survey. Thus, our target quantity is a functional of the marked point process, namely the number of unemployed people in region A, given by

$$N(A) = \sum_{j=1}^{N_1(A)} Y(s_j).$$

Let $s = (s_1, ..., s_n)$ be the location of sampling units chosen in the sampling survey, $y(s)$ the number of unemployed in each residential unit and z_1, z_2 the covariates specific to residential units and marks respectively. We denote by $\mathbf{x} = (n, s, y(s), z_1, z_2)$ the observed data obtained from the sampling survey. Our specific target quantities are the posterior predictive mean and variance of the random variable $N(A)$ given by respectively

$$\begin{aligned} \text{E}(N(A)|\mathbf{x}) &= \text{E}_{(W(s),\theta|\mathbf{x})}[\text{E}(N(A)|\mathbf{x}, W(s), \theta)] \\ &= \int_{W(s),\theta} \text{E}(N(A)|\mathbf{x}, W(s), \theta) p(W(s), \theta|\mathbf{x}) dW(s) d\theta, \end{aligned} \tag{3.6}$$

and

$$\begin{aligned} \mathrm{Var}(N(A)|\mathbf{x}) &= \mathrm{Var}_{(W(s),\theta|\mathbf{x})}\left[\mathrm{E}(N(A)|\mathbf{x}, W(s), \theta)\right] \\ &+ \mathrm{E}_{(W(s),\theta|\mathbf{x})}\left[\mathrm{Var}(N(A)|\mathbf{x}, W(s), \theta)\right]. \end{aligned} \tag{3.7}$$

Calculation of

$$\mathrm{E}(N(A)|\mathbf{x}, W(s), \theta) = \mathrm{E}(\sum_{j=1}^{N_1(A)} Y(s_j)|\mathbf{x}, W(s), \theta) \tag{3.8}$$

and

$$\mathrm{Var}(N(A)|\mathbf{x}, W(s), \theta) = \mathrm{Var}(\sum_{j=1}^{N_1(A)} Y(s_j)|\mathbf{x}, W(s), \theta), \tag{3.9}$$

require certain assumptions.

1. Conditional on $W(s)$, the point patterns of the Cox process over disjoint regions are independent. Consequently, conditional on $W(s)$, the point patterns over the design pixels I_j are also independent and we also assume that within each pixel the intensity function of the Cox process is homogeneous so that $\lambda_1(s) = \lambda_1(I_j)$ for every $s \in I_j$.

2. We assume that conditional on $W(s)$, the marks $Y(s)$ are independent of the point patterns so that the conditional intensity function of the marked point process is given by

$$\lambda(s, y|\mathbf{x}, W(s), \theta) = \lambda_1(s|\mathbf{x}, W(s), \theta) P_{Y(s)|bx,W(s),\theta}(y),$$

3. We assume that conditional on $W(s)$, marks observed on disjoint sets are independent.

4. Finally, we assume that the marks $Y(s_i)$ are identical for every $s_i \in I_j$, that is the number of unemployed in every residential unit in pixel I_j are identical. Hence, in each pixel we replace $\mathrm{E}(Y(s_i))$ by $\mathrm{E}(Y(I_j))$.

To summarize, we have two major assumptions in this model: The latent Gaussian field $W(s)$ is the only source of dependence in the model. Not only are the point patterns and marks independent conditional on $W(s)$, but the point pattern and marks are independent over disjoint intervals conditional on $W(s)$. Further, within a km^2 design unit pixels, we assume homogeneity of the point patterns as well as marks.

Let $N(I_j)$ be the number of residential units in each pixel I_j. Then with assumptions (1)-(4),

$$
\begin{aligned}
\mathrm{E}(N(A)|\mathbf{x}, W(s), \theta)) &= \mathrm{E}(\sum_{j=1}^{N_1(A)} Y(s_j)|\mathbf{x}, W(s), \theta) \\
&= \mathrm{E}(\sum_{I_j \in A} \sum_{i \in I_j} Y(s_i)|\mathbf{x}, W(s), \theta) \\
&= \sum_{I_j \in A} \mathrm{E}(\sum_{i=1}^{N(I_j)} Y(s_i)|\mathbf{x}, W(s), \theta)) \qquad (3.10) \\
&= \sum_{I_j \in A} \mathrm{E}(N_{I_j}|W(I_j))\mathrm{E}(Y(I_j)|\mathbf{x}, W(I_j), \theta) \qquad (3.11) \\
&= \sum_{I_j \in A} ||I_j||\lambda_1(I_j|\mathbf{x}, W(I_j), \theta)\lambda_2(I_j|\mathbf{x}, W(I_j), \theta) \qquad (3.12) \\
&\sim \int_{s \in A} \lambda_1(s|\mathbf{x}, W(s), \theta)\lambda_2(s|\mathbf{x}, W(s), \theta)ds \qquad (3.13)
\end{aligned}
$$

Here, $W(I_j)$ represents the latent gaussian Markov random field approximating the latent Gaussian random field $W(s)$ obtained by the SPDE method. (3.12) follows from the conditional independence and homogeneity of the point patterns as well as the marks within each km^2 pixels, whereas (3.13) follows from the approximation of integrals by sums over the design pixels. Thus the km^2 design pixels are the smallest units over which we approximate the point referenced process.

We can calculate, with similar arguments

$$
\begin{aligned}
\mathrm{Var}(N(A)|\mathbf{x}, W(s), \theta) &\sim \int_{s \in A} \lambda_1(s|\mathbf{x}, W(s)), \theta)\lambda_2(s|\mathbf{x}, W(s), \theta)ds \\
&+ \int_{s \in A} \lambda_1(s|\mathbf{x}, W(s), \theta)\lambda_2^2(s|\mathbf{x}, W(s), \theta)ds \qquad (3.14)
\end{aligned}
$$

The mean and the variance of of the predictive distribution given in (3.6) and (3.7) can be calculated numerically. INLA package permits the calculation of the intensity function $\lambda_1(s|\mathbf{x}, W(s), \theta)$ as well as the mean mark $\lambda_2(s|\mathbf{x}, W(s), \theta)$. INLA also simulates from the marginal posterior densities of the latent process as well as the model parameters, thus target quantities (3.6), (3.7) can be efficiently calculated within the INLA platform. In the next section, we briefly discuss how these calculations are carried within INLA.

3.3.2 Bayesian inference using INLA

Conditional on a realization of $W(s)$, a log-Gaussian Cox process is an inhomogeneous Poisson process. It follows that the likelihood for an LGCP is of the form

$$
\log(p(\theta|\mathbf{x})) = |\Omega| - \int_{\Omega} \lambda_1(s|\mathbf{x}, \theta)ds + \sum_{s_i \in S} \lambda_1(s_i|\mathbf{x}, \theta), \qquad (3.15)
$$

where S is the set of observed locations and $\lambda_1(s)$ is defined in (3.1).

The integral in the likelihood is intractable due the stochastic nature of $\lambda_1(s)$. To solve this problem we could use the traditional methods to fit a log-Cox process,

which consists of dividing the study regions into cells, forming a lattice, and then counting the number of points into each one. These counts are modelled using the Poisson likelihood. See for example Illian *et al* (2010). However, Simpson *et al* (2016) consider that this approach can be very inefficient, especially when the intensity of the process is high, the window of observation is too large or when the pattern is rare. They propose the use of an SPDE (Stochastic Partial Differential Equation) approach, introduced by Lindgren *et al* (2011), to transform a Gaussian field (GF) to a Gaussian Markov random field (GMRF). This methodology uses a computational mesh only for representing the latent Gaussian random field and not for modelling counts. Lindgren *et al* (2011) assume the following finite element representation

$$W(s) \approx \sum_{j=1}^{N} w_j \psi_j(s) \tag{3.16}$$

where N is the number of the mesh nodes, $w = (w_1, w_2, ..., w_N)^T$ is a multivariate Gaussian random vector (representing a Gaussian Markov random field, GMRF) and $\{\psi_j\}_{j=1}^{N}$ are the selected basis functions defined for each mesh node: ψ_j is 1 at mesh node j and 0 in all the other mesh nodes. w is chosen in a way that the distribution of $W(s)$ approximates the distribution of the solution to an SPDE. Lindgren *et al* (2011) showed that the resulting distribution for the weights is $w \sim N(0, Q(\tau, k)^{-1})$ where the precision matrix $Q(\tau, k)$ is a polynomial in the parameters τ and k. Working directly with the SPDE parameters k and τ can be difficult because they both affect the variance of the field (Yuan *et al* (2017)). So, we will consider the standard deviation σ and the spatial range ρ which are respectively given by

$$\sigma = \sqrt{\frac{1}{4\pi k^2 \tau^2}} \tag{3.17}$$

and $\rho = \frac{\sqrt{8}}{k}$. After that approximation, it follows that the integral in (3.15) can be written as

$$\int_\Omega \lambda_1(s) ds = \int_\Omega \exp(W(s)) ds \approx \int_\Omega \exp\left(\sum_{j=1}^{N} w_j \psi_j(s)\right) ds \tag{3.18}$$

This integral can be approximated using standard numerical integration schemes. Simpson *et al* (2016) suggest to use the follow quadrature rule

$$\int_\Omega f(s) ds \approx \sum_{i=1}^{N+n} \beta_i f(s_i) \tag{3.19}$$

where $\{s_i\}_{i=1}^{N+n}$ are the locations of mesh nodes and observations, and $\{\beta_i\}_{i=1}^{N+n}$ are the quadrature weights.

Unlike the traditional methods for inference in LGCP models, this methodology uses each location to model the point pattern, without aggregation. The LGCP model belongs to the latent Gaussian models and consequently, the inference can be done within the INLA platform.

As we noted, the SPDE methodology requires a triangulation of the study region to represent a Gaussian random field. Here, we used a Delaunay triangulation with 3923 mesh nodes.

In real data applications, it is common that the point pattern and its associated marks are dependent. In our case, we expect that the average number of unemployed people per dwelling to be dependent with the intensity of residential buildings, but the sign of that correlation is not obvious. On the one hand, we expect the number of unemployed people to be higher in regions with higher intensity of residential buildings and on the other hand, we expected more opportunities of employment in these regions.

Illian *et al* (2008) describes two types of marked point process models depending on the type of dependence between the point patterns and marks. Here, we consider two versions of conditional dependence: In the first model we assume, as was explained in section 3.3, that there is a common latent Gaussian field that govern the dependence structures of points and marks and conditional on this field, point patterns and marks are independent. In the second alternative model, we assume that there are two independent fields that govern the dependence structures of points patterns and marks. It is also possible to introduce a third coreginalization model (Banerjee *et al*, 2004, Gelfand *et al*, 2004) consisting of two dependent latent processes for point patterns and marks by assuming independence of points and marks conditional on these latent processes. Coreginalization models can be inferred within the INLA platform, however this would require joint modelling of two dependent fields and we will not pursue these models in this work. In table 4.1, a comparison of these alternative models is given.

Here, we give the details of the model based on a common latent Gaussian model. Let us consider that $\{s_i\}_{i=1}^{N+n}$ are the locations of the mesh nodes and the locations of the sampled residential buildings, and $\{y(s_i)\}_{i=1}^{N+n}$ are the number of unemployed people per residential building. The hierarchical structure of the model considered is given by

1. Data|Parameter

$$p(\{s_i, i = 1, ..., N + n\}|\lambda_1) \approx \prod_{i=1}^{N+n} \text{Poisson}(\beta_i \lambda_1(s_i)) \tag{3.20}$$

$$p(\{y(s_i), i = 1, ..., N + n\}|\lambda_2) \approx \prod_{i=1}^{N+n} \text{Poisson}(\lambda_2(s_i)) \tag{3.21}$$

   where β_i is defined in 3.19.

2. Parameter|Hyperparameters

$$\log(\lambda_1(s_i)) = \alpha_1 + \text{offset}_1(s_i) + W(s_i), \tag{3.22}$$

$$\log(\lambda_2(s_i)) = \alpha_2 + \text{offset}_2(s_i) + Z_2'(s_i)\theta + \alpha_3 W(s_i), \tag{3.23}$$

   where $W(s)$ is the GMRF given in (3).

3. Hyperparameters

$$\alpha_1 \sim N(0, 1000) \tag{3.24}$$

$$\alpha_2 \sim N(0, 1000) \tag{3.25}$$

$$\theta_j \sim N(0, 1000), j = 1, ..., p \tag{3.26}$$

$$\alpha_3 \sim N(0, 1000) \tag{3.27}$$

We assume that the latent field W belong to the Matern class with $\nu = 1$. We further assume that the model parameter of this field has the same prior structure as given below:

We followed Simpson *et al* (2017) and Fuglstad *et al* (2017) to construct a joint penalising complexity (PC) prior density for the spatial range, ρ, and the marginal standard deviation, σ, which is given by

$$p(\rho, \sigma) = RS\rho^{-2}e^{-R\rho^{-1}-S\sigma} \tag{3.28}$$

where R and S are hyperparameters determined by $R = -\log(\alpha_1)\rho_0$ and $S = \frac{-\log(\alpha_2)}{\sigma_0}$.

The practical approach for this in INLA is to require the user to indirectly specify these hyperparameters through $P(\rho < \rho_0) = \alpha_1$ and $P(\sigma < \sigma_0) = \alpha_2$. Here, we considered $\rho_0 = 400, \alpha_1 = 0.5, \sigma_0 = 1, \alpha_2 = 0.5$.

The term $\text{offset}_1(s_i)$ in (3.22) represents the log population density. We know their numbers by NUTS III regions so, based on that, we produced a spatial prediction for all domains by way of a Kernel smoothing (explained in the section 1.5.2), using the centroids of the NUTS III regions. The resulting prediction is given in figure 3.1. Grande Lisboa, Grande Porto and Península de Setúbal are the regions that stand out most. The $\text{offset}_2(s_i)$ term in (3.23) represents the log of the number of people per residential building. We have the information for the residential buildings locations in the sample, but we need to estimate it for the mesh nodes. For this we also used the Kernel smoothing.

Model selection

To evaluate the significance of each covariate and random effect in the marks, we considered different models and compared the results of two model selection criteria: deviance information criterion (DIC) and Watanabe-Akaike information criterion ($WAIC$).

DIC, proposed by Spiegehalter *et al* (2002), is the most commonly used measure of model fit. It is based on a balance between the fit of the model to the data and the corresponding complexity of the model: $DIC = \bar{D} + p_D$ where $\bar{D}$ is the posterior mean deviance of the model and p_D is the effective number of parameters. The model with the smallest value of DIC is the one with a better balance between the model adjustment and complexity. However, this criterion can present some problems, which arise in part from not being fully Bayesian.

A typical alternative is the WAIC, proposed by Watanabe (2010), which is fully Bayesian in that it uses the entire posterior distribution. It can be considered as an improvement on the DIC for Bayesian models (Gelman *et al.*, 2014).

Several alternative spatial random effects were used in modelling the intensities, namely (i) Common random effect W both for points and marks (ii) Random effect W and its scaled version $\alpha_3 W$ for the points and marks respectively (iii) two independent latent processes W_1 and W_2 for the points and their marks respectively.

Table 3.1 shows the values of these two criterions for the models considered for the marked point process. In this case, the model with the best performance was the one that took into account the following factors:

- the offset term given by the population density to model the intensity of the point process (offset_1) ;
- the covariates to model the mark intensity (number of individuals per residential building($nind_2$): the median of the education level (edu_2), the mean age (age_2), and the proportion of registered unemployed people (iefp_2). Here, subscripts 1 and 2 indicate that the corresponding covariate is used in modelling intensity $\lambda_1(s)$ and $\lambda_2(s)$ respectively;
- two independent latent processes W_1 and W_2 used for points and their marks.

log intensity of points and marks	DIC	$WAIC$	p_{DIC}	p_{WAIC}
α_1 ; α_2	110707.09	110724.00	2.18	19.04
$\alpha_1 + offset_1$; α_2	84826.76	84853.04	2.198	28.41
$\alpha_1 + offset_1$; $\alpha_2 + offset_2$	110707.09	110724.00	2.18	19.04
$\alpha_1 + offset_1$; $\alpha_2 + nind_2$	84791.91	84818.07	3.199	29.29
$\alpha_1 + offset_1$; $\alpha_2 + nind_2 + edu_2$	84790.11	84816.27	4.198	30.28
$\alpha_1 + offset_1$; $\alpha_2 + nind_2 + edu_2 + age_2$	84714.83	84741.04	5.198	31.34
$\alpha_1 + offset_1$; $\alpha_2 + nind_2 + edu_2 + age_2 + iefp_2$	84706.69	84733.00	6.197	32.444
$\alpha_1 + offset_1 + W$; $\alpha_2 + nind_2 + edu_2 + age_2 + iefp_2$	48538.24	67474.12	1817.38	15031.22
$\alpha_1 + offset_1 + W$; $\alpha_2 + nind_2 + edu_2 + age_2 + iefp_2 + \alpha_3 W$	48656.62	67562.53	1796.70	14998.79
$\alpha_1 + offset_1 + W_1$; $\alpha_2 + nind_2 + edu_2 + age_2 + iefp_2 + W_2$	48505.25	67443.79	1842.59	15058.96

Table 3.1: DIC, WAIC and the effective number of parameters

Here, p_{DIC} and p_{WAIC} are the effective number of parameters, as described in Spiegelhalter *et al* (2002) and Gelman *et al* (2014), respectively.

It is clear from the table that the model that employs all the covariate information and two independent latent processes, one for points other for marks seems to give the best fit with the model that employs all the covariate information and a single common latent process for the points and marks coming second. Here, we chose the model with lower DIC to continue with these analysis. We also considered a negative binomial distribution for the marks as an alternative to the poisson marks, but these models did not bring gain in terms of DIC.

To perform the spatial prediction, we created a regular grid of 1km^2 in the domain. A projection from the mesh to the grid was performed and the resultant maps of the posterior mean of the logarithmic transformation of the intensity of the residential units $\log(\lambda_1(s))$ and the logarithmic of the marks mean are shown

in figure 3.2. The plot of the logarithmic transformation of the intensity provides a clearer image about the spatial variation of the residential buildings. As we expected, the highest values are concentrated in Grande Lisboa, Grande Porto, and Algarve regions. The intensity is clearly higher near the coast and lower in the interior of the country.

The standard deviations of these fields are plotted in figure 3.3.

With these estimates, we can conclude that the average number of unemployed people per residential building is higher in the Grande Porto, Península de Setúbal and Alentejo Central regions.

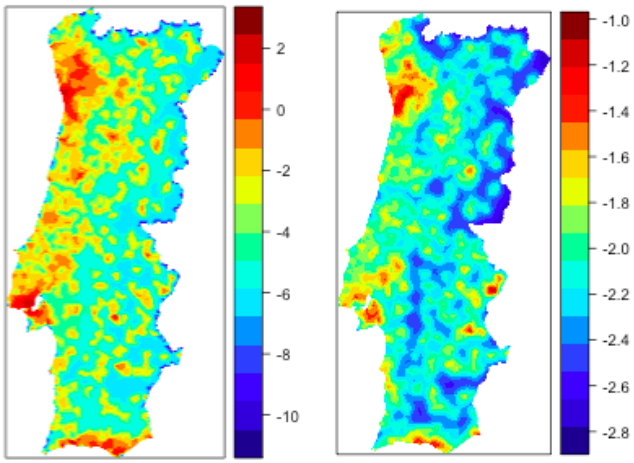


Figure 3.2: Posterior mean of: log intensity (left), and log mean marks (right)

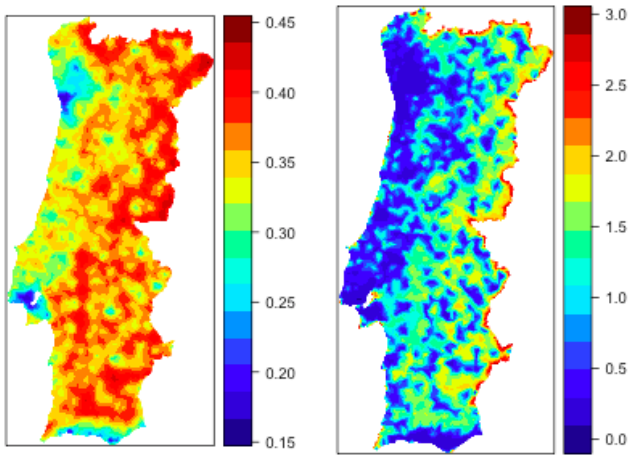


Figure 3.3: Standard deviation of: log intensity (left), and log mean marks (right)

3.3.3 Model validation

For model validation purposes, we chose randomly 26 municipalities and fitted the model excluding the data from these municipalities. Figure 3.4 gives the 95% credible intervals for the predicted values of unemployment from the model together

with the observations and predicted values for these 26 municipalities. Figure 3.5 gives the credible intervals together with observations and estimates for the same 26 chosen municipalities when all the data are used in fitting the model. Figure 3.6 gives the Pearson residuals versus fitted values for the 278 municipalities. Figures 3.7 gives the 95% credible intervals, observations and estimates for the NUTS III regions. As is expected, the model gives higher precision estimates at NUTS III regions as compared to municipality level.

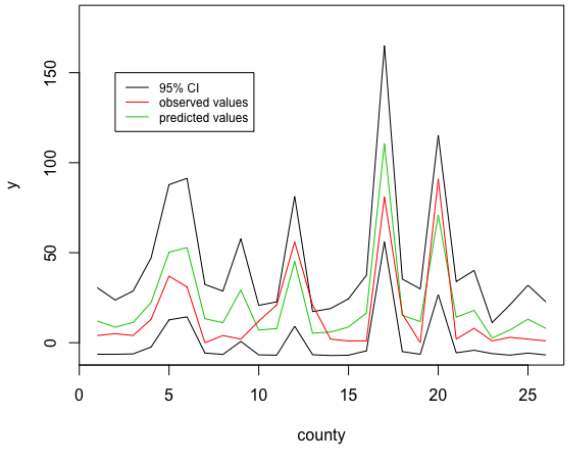


Figure 3.4: 95% CI, observed values and predicted values for the 26 municipalities that were removed from the sample

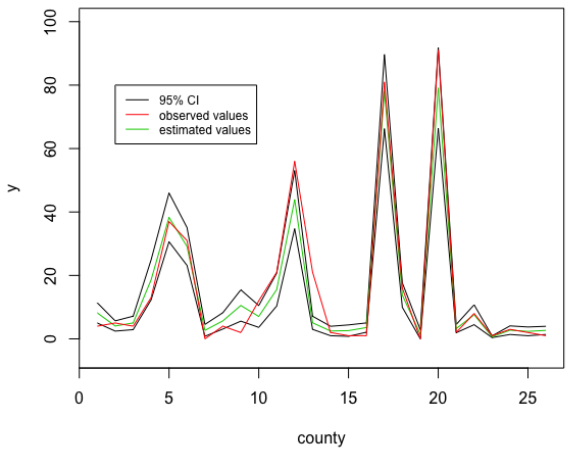


Figure 3.5: 95% CI, observed values and predicted values for the 26 municipalities that were removed from the sample for the first plot. Here, we used all municipalities in the sample for the modelling process.

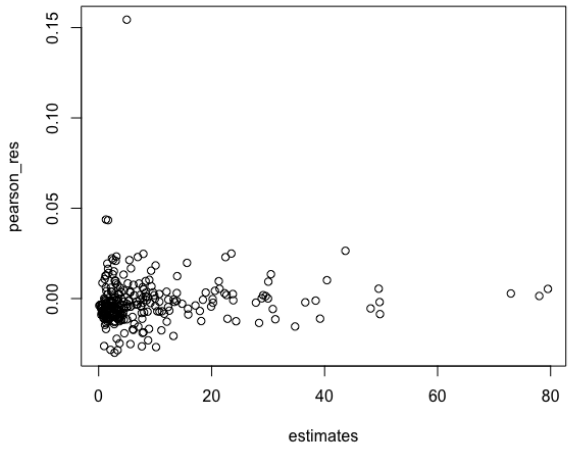


Figure 3.6: Pearson residuals versus fitted values for the 278 municipalities

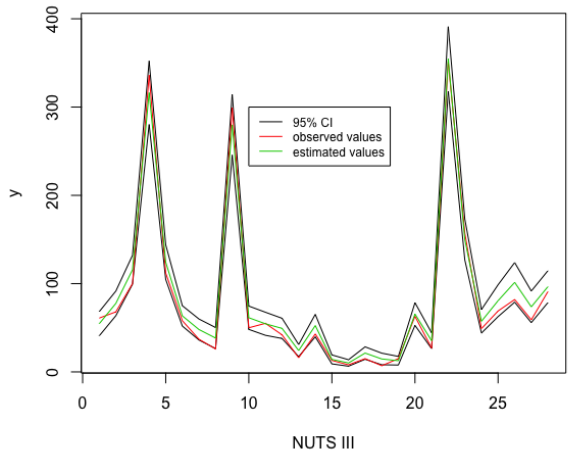


Figure 3.7: 95% CI, observed values and fitted values for the 28 NUTS III regions

3.3.4 Unemployment estimation

The marked point process model explained in the previous section, projects the sampling survey in space. This point process is a thinned version of the true point patterns of the residential units together with their marks across Portugal.

Let $N_1^*(s)$ represent the true point patterns of the residential units with intensity $\lambda_1^*(s)$. Then

$$\lambda_1^*(s) = \frac{\lambda_1(s)}{P(RU(s))}, \tag{3.29}$$

where, $P(RU(s))$ is the probability that a residential unit at s is included in the survey. $P(RU(s))$ should be interpreted as the proportion of the residential units in any infinitesimal area which is included in the sampling survey.

Assume also that $N_2^*(s)$ represent the true intensity of the number of unemployed observed in residential unit at location s. then the intensity $\lambda_2^*(s)$ of this counting process is given by

$$\lambda_2^*(s) = \frac{\lambda_2(s)}{P(D(s)|RU(s))}, \tag{3.30}$$

where the probability $P(D(s)|RU(s))$ should be interpreted as the proportion of dwellings in a residential unit which are included in the sampling survey.

Target quantities 3.6 and 3.7 depend on the multiplicative intensity $\lambda_1(s)\lambda_2(s)$, which is a thinned version of $\lambda_1^*(s)\lambda_2^*(s)$ and this relationship is given by

$$\lambda_1^*(s)\lambda_2^*(s) = \frac{\lambda_1(s)\lambda_2(s)}{p(s)}, \tag{3.31}$$

where

$$\begin{aligned} p(s) &= P(RU(s))P(D(s)|RU(s)) \\ &= P(D(s)), \end{aligned}$$

since $P(D(s)|RU^c(s)) = 0$.

Here, $p(s)$ should be interpreted as the proportion of dwellings that are chosen in the sampling survey. As explained in section 1.2 these probabilities are estimated using (1.1).

To define the intensity of the full version of the spatial point process, the knowledge of the sampling probabilities $p(s)$ for whole domain is required. Here, we estimate these probabilities using the kernel method, based on the data given by the sampling survey. This method allows us to generate a spatial prediction for the centers of the cells of the grid, derived from the values of the dwellingss locations.

We simulated 1000 values of the predictive posterior distribution of λ_1 and λ_2 for each cell I_j to estimate the target quantities, by simulating samples from the posterior distributions of the model parameters and the latent gaussian markov fields used in the model.

Figure 3.8 gives the predictive multiplicative intensity function

$$\frac{\lambda_1(s|\mathbf{x})\lambda_2(s|\mathbf{x})}{p(s)}, \tag{3.32}$$

which will form the basis for calculating the unemployment in any region A, expressed in terms of $E(N(A)|\mathbf{x})$ as given in 3.6.

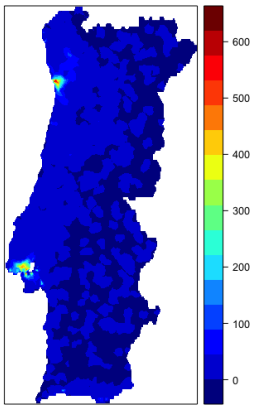


Figure 3.8: Predictive multiplicative intensity function

We calculated the credible intervals for the posterior mean by NUTS III and compared those with the direct estimates (figure 3.9). We note that the direct estimates are inside the intervals for almost all regions. The highest points correspond to Grande Porto and Grande Lisboa regions. The figure reveals that there is an underestimation in these regions and an overestimation in the others. This behaviour is probably due to the large differences in the intensity in these regions and consequent smoothing of intensities across the space.

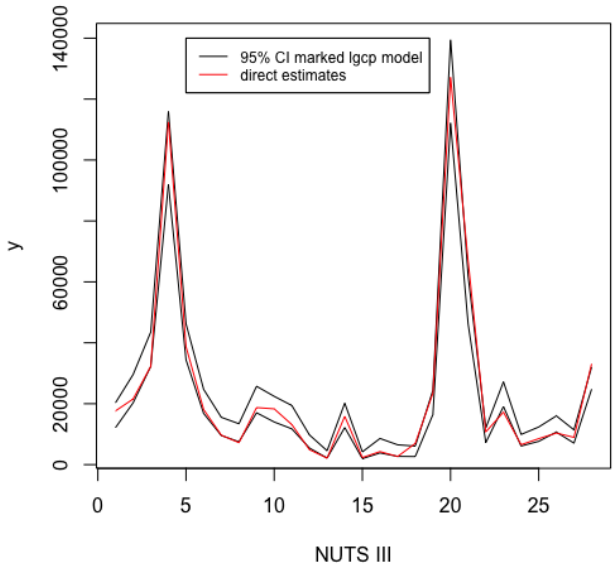


Figure 3.9: 95% credible intervals for the posterior mean (black) and direct estimates (red) by NUTS III regions

3.4 Comparison between the results of the marked LGCP model and the traditional small area models

We conduct a comparison between the marked LGCP model described in the previous section, the direct method and some traditional small area models. For comparison, we choose the hierarchical Bayesian model versions of the standard Fay and Herriot model (FH model) proposed by Rao and Molina (2015) and its extension with a latent CAR model (FH-CAR model) to borrow strength from adjacent areal units given by You *et al* (2011). In both of these models, we use the same set of auxiliary information, aggregated to NUTS III regions and we report the total unemployed estimates for 28 NUTS III regions obtained from these 3 alternative methods. We also report the performance of these 3 models on higher resolution, namely the estimates of total unemployed in 278 municipalities.

Figure 3.10 shows the estimates obtained by the direct method, the FH model, the FH-CAR model and the LGCP model. We can see that the FH models (FH and FH-CAR) disagree with the direct method in the Lisbon region. This result, in the most densely populated regions, may indicate some fragility. Moreover, coefficients of variation (CVs), given by the ratio of the standard deviation to the mean, obtained by the FH model for the Serra da Estrela region are quite high (figure 3.11).

As we expected, for the majority of the regions the FH model presents lower CVs in comparison with the FH-CAR model and the LGCP model. This result can be explained by the difference in the number of parameters in these models. Moreover, the FH models use the direct estimates as the data in the modelling process and assume that the variances are fixed. The LGCP model uses the data from the sample of the Labour Force Survey, and do not assume the variances as fixed. In any case, the CVs of the estimates obtained by the FH-CAR models, as well as the CVs of the estimates obtained by the LGCP model do not achieve 25% in any NUTS III region.

In general, the FH-CAR and LGCP models present similar CVs. Since the estimates and the CVs of these two models are close, we think that the LGCP model here proposed brings many advantages to this problem, as we explained in the introduction section.

Notice that if the direct method is applied to a more disaggregated geographical level such as municipalities, it is not possible to provide estimates for some regions (figure 3.12. That regions are regions without observations in the LFS or without any unemployment observation. Consequently, the basic FH models can not provide estimates for that regions since they use direct estimators as data with known variances. The FH-CAR model can do that using the spatial effects. However, in some municipalities there are not observations even in the neighboors. Thus, the estimates in that municipalities are questionable. Figure 3.12 shows that in the majority of regions with observations in the LFS, the LGCP model is the one with lowest CVs. For a clear comparison, we show the boxplots of the CVs obtained from each model in figure 3.13. As we expected, the higher the level of disaggregation, the greater the difference between the LGCP and the traditional SAE models in

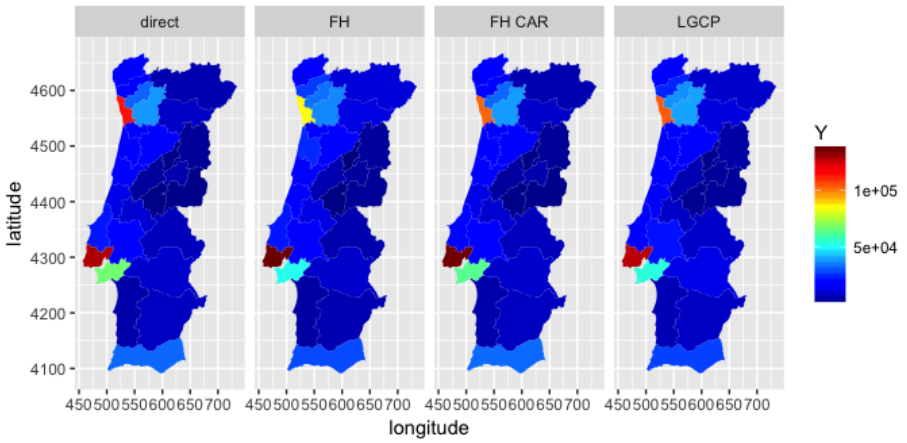


Figure 3.10: Estimates of total unemployed from the direct method, FH model, FH-CAR model and LGCP model, by NUTS III for the 4th quarter of 2014

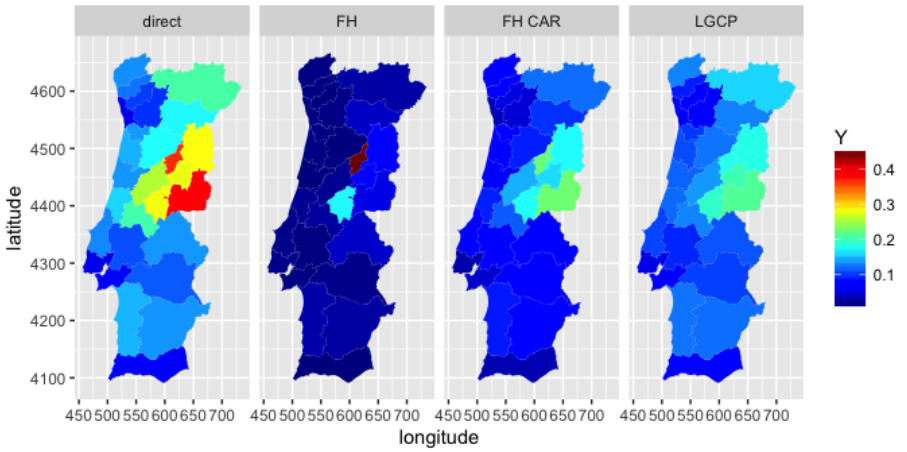


Figure 3.11: Coefficients of variation of the estimates obtained by the direct method, FH model, FH-CAR model and LGCP model

terms of variability.

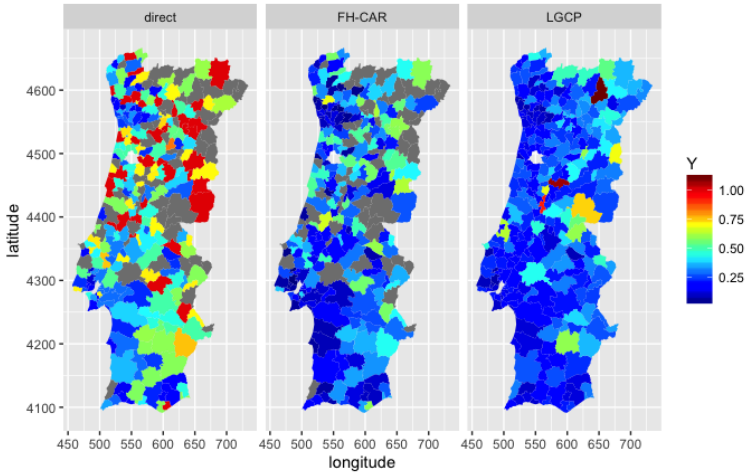


Figure 3.12: CVs of the estimates obtained by the direct method (left), FH-CAR model (middle) and LGCP model (right)

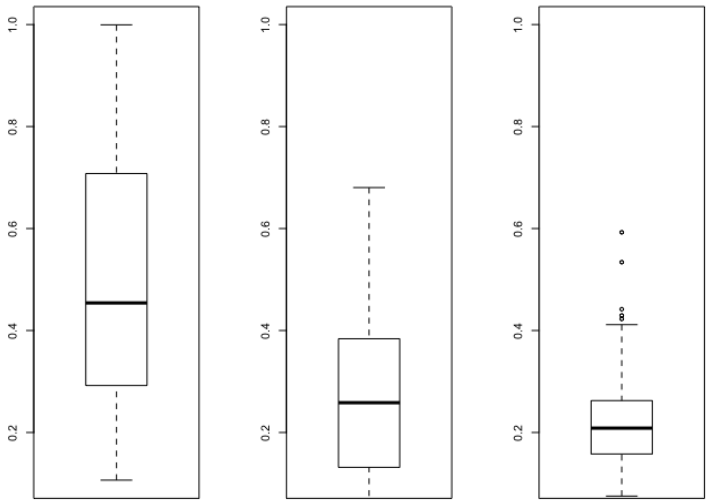


Figure 3.13: Boxplots with the CVs of the estimates obtained by the direct method (left), FH-CAR model (middle) and LGCP model (right)

3.5 Discussion

In this study, we employed point referenced spatial models to unemployment estimation making use of the newly available geo-referenced locations of sampling units. Unlike the direct method and area level models, our approach take into account this

important information, permitting spatial smoothing at higher spatial resolutions, making use of the disaggregated data.

Although for large areal units such as NUTS III, unemployment figures estimated from our method do not differ significantly from the estimates obtained from the standard areal models employed within the SAE methodology, main benefits of the proposed method can clearly be seen as one seeks for unemployment estimates at smaller areal units, such as municipalities. We did not carry out a full sensitivity analysis, however, this analysis, carried along the lines suggested by Roos *et al*(2015) may suggest how the proposed models can be improved.

National Statistical Institutes usually require a great deal of consistency between the estimates obtained at different areal resolutions, but this requirement is not easy to satisfy using the standard small area models. Since this new methodology operates independently from the administrative limits of geographical units, it can provide us with the necessary means to meet this requirement both a consistent and timely fashion.

Despite the relative complexity of this methodology, the computational costs are not high due to the availability of the R-INLA package in the software R.

In this work, we do not report on the time dynamics of unemployment. At present, there are only 14 quarterly sample surveys with geo-referenced sampling units. As these quarterly data further become available, it may be possible to investigate, with certain degree of precision, how spatial variation of unemployment changes over time. This can be done by considering space-time marked point processes, in which the latent process now is a space time Gaussian process. By adding time varying covariates in the model such as a linear or quadratic trends functions in time, it may be possible to capture, in detail, the time dynamics of the unemployment across the domain of study. It is possible to infer on such models within the INLA platform with some moderate increase on computational time.

Recently, NSI has started on a much more ambitious geo-referencing method by identifying and geo-referencing not only the sampled residential units, but all the residential units in Portugal. Point level methods and models that are adequate for these new realities will be discussed elsewhere as new data become available.

Chapter 4

Geostatistics modelling

4.1 Introduction

Recently, NSI extended and improved their data gathering methods by georeferencing every residential unit across Portugal. With this new detailed georeferencing method, spatial distributions of residential units are no longer random. Therefore, new spatial models without the need to model randomness of points should, in principle, produce estimates which are more precise and with reduced sampling variation. Hence, with such new information, the objective becomes to model the spatial variation of the marks using point referenced methods, and then to extrapolate this in space to all georeferenced residential units.

In addition to the extrapolation in space, we intend to do a temporal extrapolation. The temporal extension will be based on 9 sequentially observed quarterly sampling surveys (from the 4th quarter of 2014 to the 4th quarter of 2016).

For the modelling process, we suggest using a geostatistical model with a temporally and spatially structured random effect. Typically, in this framework, the spatial process is a Gaussian field (GF). Inference on such models is not straightforward due to the dense covariance matrices, a problem known in the literature as *big n problem* (Banerjee et al, 2004). Due to the computational problems that emerge in this framework, Lindgren *et al* (2011) proposed a more computationally tractable approach based on stochastic partial differential equation (SPDE) models. These models allow for the transformation of a Gaussian field into a Gaussian markov random field, and as a result we follow this method.

4.2 Data

In each quarter, the met sample size (respondents) is around 35000 observations, distributed across 14000 dwellings, located in roughly 13800 residential buildings. Therefore, in the majority of the sampled buildings, only one dwelling is selected. Each individual in the sample is interviewed about their status in the labour market (employed, unemployed, inactive), sex, age and education level (primary level, secondary level, higher level), etc.

The georeferencing of all residential buildings is available, even those units that are not included in the sample. Although a residential building can consist of multi-

ple dwellings, the geographical coordinates are available only for the buildings themselves. Consequently, and particularly in areas of high population density, multiple dwellings in the survey may have the same spatial location. Hence, to avoid an overlap in the locations within the modelling process, the observation units we will consider are the residential buildings. In the following sections, we will denote the average number of unemployed people per dwelling in the residential building at s_j location and quarter t by $y(s_j, t)$ (rounding to the nearest integer). Here, we intend to extrapolate the values observed in the sampled locations to all residential buildings with at least one dwelling as usual residence (around 2300000). The number of dwellings per residential building is known. Note that 5/6 of the sampled dwellings are unchanged from one quarter to another due to the rotative sampling design explained in the previous section. In general, each individual is interviewed in 6 consecutive quarters, which results in a high temporal correlation in our data.

In the modelling process, we use some covariates at residential building level for each quarter, namely the mean age and the median of the education level. Although the education level does not constitute a quantitative variable, it was treated as such due to its ordinal meaning (1-primary level, 2-secondary level, 3-higher level). The average number of people per dwelling in each residential building was considered as an offset. We also use information about the proportion of unemployed people registered in the employment centers, available by municipalities. A spatial extrapolation of the covariates for the whole domain and study period is required (figure 4.1 shows the covariates for the 4th quarter of 2016). For this, we used a Kernel method.

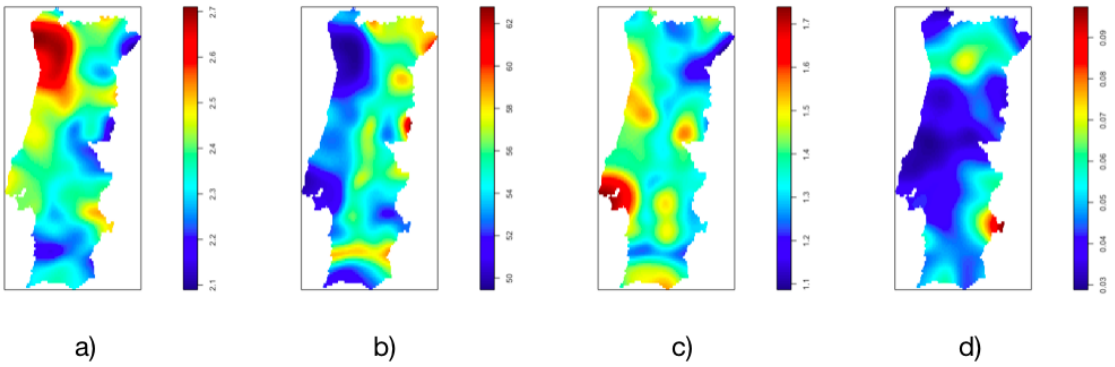


Figure 4.1: Kernel estimates of: a) the average number of people per dwelling in each residential building; b) mean age per residential building; c) median of the education level per residential building; d) proportion of registered unemployed people in centers of employment

4.3 Bayesian models for point referenced data

We will assume a Poisson distribution for $y(s, t)$:

$$y(s,t)|\lambda(s,t) \sim \text{Poisson}(\lambda(s,t)) \tag{4.1}$$

with

$$\log(\lambda(s,t))|W(s,t) = \alpha + \text{offset}(s,t) + \sum_{m=1}^{M} \theta_m z_m(s,t) + W(s,t), \tag{4.2}$$

where $W(s,t)$ is a latent spatio-temporal process, $\theta = c(\alpha, \{\theta_m, m = 1, ..., M\})$ are the model parameters, and $\{z_m(s,t), m = 1, ..., M\}$ are the covariates.

4.3.1 Target quantities for inference

The total number of unemployed people in a given area A and quarter t is given by

$$N(A,t) = \sum_{s_j \in A} y(s_j,t) N(s_j) \tag{4.3}$$

where $N(s_j)$ is the number of dwellings in the residential building at s_j (known quantity).

We also note that contrary to the point process model, for any region A, the number of s_j in A are also known and fixed within this new georeferencing scheme.

We denote by $\mathbf{x}(t) = (s_j, t, y(s_j,t), z(s_j,t))$ the observed data obtained from the sampling survey in quarter t, and $\mathbf{x} = (\mathbf{x}(1), ..., \mathbf{x}(9))$ the collected data in the 9 quarters of study.

Our specific target quantities are the posterior predictive mean and variance of the random variable $N(A,t)$ given respectively by

$$\begin{aligned} \mathrm{E}(N(A,t)|\mathbf{x}) &= \mathrm{E}_{(W(s,t),\theta|\mathbf{x})}[\mathrm{E}(N(A,t)|\mathbf{x}, W(s,t), \theta)] \\ &= \int_{W(s,t),\theta} \mathrm{E}(N(A,t)|\mathbf{x}, W(s,t), \theta) p(W(s,t), \theta|\mathbf{x}) dW(s,t) d\theta \end{aligned} \tag{4.4}$$

and

$$\begin{aligned} \mathrm{Var}(N(A,t)|\mathbf{x}) &= \mathrm{Var}_{(W(s,t),\theta|\mathbf{x})}\left[\mathrm{E}(N(A,t)|\mathbf{x}, W(s,t), \theta)\right] \\ &+ \mathrm{E}_{(W(s,t),\theta|\mathbf{x})}\left[\mathrm{Var}(N(A,t)|\mathbf{x}, W(s,t), \theta)\right]. \end{aligned} \tag{4.5}$$

Calculation of

$$\mathrm{E}(N(A,t)|\mathbf{x}, W(s,t), \theta) = \mathrm{E}(\sum_{s_j \in A} y(s_j,t) N(s_j)|\mathbf{x}, W(s,t), \theta) \tag{4.6}$$

and

$$\mathrm{Var}(N(A,t)|\mathbf{x}, W(s), \theta) = \mathrm{Var}(\sum_{s_j \in A} y(s_j,t) N(s_j)|\mathbf{x}, W(s,t), \theta), \tag{4.7}$$

require certain assumptions.

1. We assume that conditional on $W(s,t)$, the observations on disjointed sets are independent.

2. We divided our spatial domain in pixels of $1km^2$ denoted by I_j (similar to the grid INSPIRE explained in the sampling design section, except in the border, which was simplified). Furthermore, we assume that the $y(s_i, t)$ are identical for every $s_i \in I_j$ in quarter t, and that the number of unemployed in every residential unit in pixel I_j in quarter t are identical. Hence, in each pixel we replace $\mathrm{E}(y(s_i, t))$ by $\mathrm{E}(y(I_j, t))$.

3. We also assume that during the study period, the number of dwellings in each pixel I_j, $N(I_j)$, does not change with time.

Then

$$\begin{aligned}
\mathrm{E}(N(A,t)|\mathbf{x}, W(s,t), \theta)) &= \mathrm{E}(\sum_{s_j \in A} y(s_j,t)N(s_j)|\mathbf{x}, W(s,t), \theta) \\
&= \mathrm{E}(\sum_{I_j \in A} \sum_{s_i \in I_j} y(s_i,t)N(s_i)|\mathbf{x}, W(s,t), \theta) \\
&= \sum_{I_j \in A} \mathrm{E}(\sum_{s_i \in I_j} y(s_i,t)N(s_i)|\mathbf{x}, W(s,t), \theta) \qquad (4.8) \\
&= \sum_{I_j \in A} N(I_j)\mathrm{E}(y(I_j,t)|\mathbf{x}, W(I_j,t), \theta) \qquad (4.9) \\
&= \sum_{I_j \in A} N(I_j)\lambda(I_j,t|\mathbf{x}, W(I_j,t), \theta) \qquad (4.10) \\
& \qquad (4.11)
\end{aligned}$$

where $W(I_j, t)$ represents the latent Gaussian Markov random field approximating the latent Gaussian random field $W(s, t)$ obtained by the SPDE method. $N(I_j)$ is calculated by counting all the dwellings in each grid cell.

The variance $\mathrm{Var}(N(A,t)|\mathbf{x}, W(s,t), \theta)$ is calculated in a similar way, and is equal to $\mathrm{E}(N(A,t)|\mathbf{x}, W(s,t), \theta))$.

The mean and the variance of the predictive distribution given in (4.4) and (4.5) can be calculated numerically. We used 1000 samples from an approximated posterior of the fitted model, calculated in the INLA platform. In the next section, we briefly discuss how these calculations are performed.

4.3.2 Bayesian inference using INLA

As we introduced in the previous section, we will assume the following hierarchical model

1. Data|Parameter
$$y(s,t)|\lambda(s,t) \sim \mathrm{Poisson}(\lambda(s,t)) \qquad (4.12)$$

2. Parameter|Hyperparameters
$$\log(\lambda(s,t))|W(s,t) = \alpha + \mathrm{offset}(s,t) + \sum_{m=1}^{M} \theta_m Z_m(s,t) + W(s,t), \qquad (4.13)$$

where $W(s,t)$ is the latent spatio-temporal process, as described in Blangiardo *et al.* (2015):

$$W(s,t) = aW(s,t-1) + \xi(s,t) \tag{4.14}$$

with $t = 1, ..., 9$, $|a| < 1$, and $W(s,1) \sim Normal(0, \sigma^2/(1-a^2))$. The term $\xi(s,t)$ is a Gaussian field with mean zero, temporally independent and with the following covariance function

$$cov(\xi(s,t), \xi(j,u)) = \begin{cases} 0, & if\ t \neq u, \\ cov(\xi(s), \xi(j)), & if\ t = u. \end{cases}$$

3. Hyperparameters

$$\alpha \sim N(0, 1000) \tag{4.15}$$

$$\theta_m \sim N(0, 1000),\ m = 1, ..., M \tag{4.16}$$

We assume that the latent field ξ belongs to the Matern class with $\nu = 1$.

The traditional modelling approach involves a Cholesky factorization of the covariance matrix. Since that matrix is dense, the operation is of the order $O(n^3)$, where n is the number of locations where the process is observed. When n is large, the process has high computational costs, and this problem is known as 'big n problem'.

An alternative approach is to approximate the Gaussian field (GF) ξ by a Gaussian Markov random field (GMRF), which is a discretized representation. That approximation is based on the stochastic partial differential equation (SPDE) approach (see Lindgren et al, 2011), and depends on a triangulation, known as a mesh, of the spatial domain. Figure 4.2 shows the mesh we considered in this study. Using GMRF models, the computational costs of working with a sparse precision matrix is of the order $O(n^{3/2})$, which is a significant difference in comparison with the operations with a dense covariance matrix. The GMRF representation of the GF allows us to use the INLA approach. Notice that the implementation of this model in OpenBUGS, without some approximations and using the sparsity of the precision matrix, would not be possible.

Lindgren *et al* (2011) assume the following finite element representation

$$\xi(s) \approx \sum_{j=1}^{N} \tilde{\xi}_j \phi_j(s) \tag{4.17}$$

where N is the number of the mesh nodes, $\tilde{\xi} = (\tilde{\xi}_1, \tilde{\xi}_2, ..., \tilde{\xi}_N)^T$ is a multivariate Gaussian random vector (representing a Gaussian Markov random field, GMRF) and $\{\phi_j\}_{j=1}^N$ are the selected basis functions defined for each mesh node: ϕ_j is 1 at mesh node j and 0 in all the other mesh nodes. ξ is chosen in a way that the distribution of $\xi(s)$ approximates the distribution of the solution to an SPDE. Lindgren *et al* (2011) showed that the resulting distribution for the weights is $\xi \sim N(0, Q(\tau,k)^{-1})$ where the precision matrix $Q(\tau,k)$ is a polynomial in the parameters τ and k. Working directly with the SPDE

parameters k and τ can be difficult because they both affect the variance of the field (Yuan *et al* (2017)). So, we will consider the standard deviation σ and the spatial range ρ which are given respectively by $\sigma = \sqrt{\frac{1}{4\pi k^2 \tau^2}}$ and $\rho = \frac{\sqrt{8}}{k}$.

Here, we use the temporal extension given by

$$\xi(s,t) \approx \sum_{j=1}^{N} \tilde{\xi}_j \phi_j(s,t) \tag{4.18}$$

We followed Simpson *et al* (2017) and Fuglstad *et al* (2017) to construct a joint penalising complexity (PC) prior density for the spatial range, ρ, and the marginal standard deviation, σ, which is given by

$$p(\rho,\sigma) = RS\rho^{-2}e^{-R\rho^{-1}-S\sigma} \tag{4.19}$$

where R and S are hyperparameters determined by $R = -\log(\alpha_1)\rho_0$ and $S = \frac{-\log(\alpha_2)}{\sigma_0}$.

The practical approach for this in INLA is to require the user to indirectly specify these hyperparameters through $P(\rho < \rho_0) = \alpha_1$ and $P(\sigma < \sigma_0) = \alpha_2$. Here, we considered $\rho_0 = 400, \alpha_1 = 0.5, \sigma_0 = 1, \alpha_2 = 0.5$.

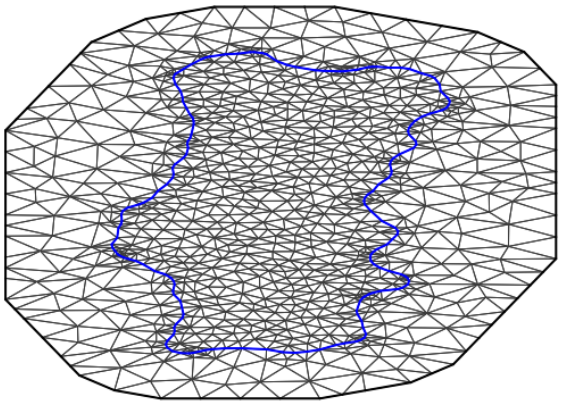


Figure 4.2: mesh with 913 vertices

Since it was necessary to know the covariates and offset in the locations of the observations and in those of the mesh nodes, we predicted them using a kernel estimation method, as explained in section 4.2 (see figure 4.1).

The inference was made using the INLA approach (Rue et al, 2009 and Rue et al, 2017).

Conducting a direct spatial prediction for the non-sampled residential buildings would be computationally expensive due to the high number of locations (around 2,300,000). For this reason, we perform the spatial prediction for the grid locations and assume that in each grid cell of $1km^2$ the number of unemployed people per dwelling is the same. To make that prediction, we follow the work of Blangiardo et al (2015) and project the latent field (estimated at the mesh nodes) onto the grid

locations, using a projector matrix A with entries $A_{ij} = \phi_j(s_i)$. This matrix is the link between the spatial latent field defined on a mesh, and the observations defined in a set of locations. In our mesh, each spatial location is placed inside a triangle which is limited by three vertices. Thus, matrix A includes three non-zero elements for each row whose sum is equal to 1. Computational details can be seen in the appendix. With this approach, the computational time is about 25 minutes using the Rue's server, and about 50 minutes using a common computer.

4.3.3 Model selection

For the model selection we used the deviance information criterion (DIC) and Watanabe-Akaike information criterion (WAIC), proposed by Spiegehalter *et al.* (2002) and Watanabe (2010) respectively. Table 4.1 shows that the best model is the one that considers the offset, the age, the proportion of unemployed people registered in the employment centers and the spatio-temporal random effects.

model	DIC	$WAIC$	p_{DIC}	p_{WAIC}
α	105324.06	105324.17	1.16	1.28
α + offset	105120.28	105120.39	1.16	1.28
α + offset + age	104799.76	104800.01	2.16	2.42
α + offset + age + edu	104801.57	104801.89	3.16	3.49
α + offset + age + $IEFP$	104459.93	104460.23	3.16	3.46
α + offset + age + $IEFP$ + W	103550.23	103569.02	306.41	324.02

Table 4.1: DIC, WAIC and the effective number of parameters

4.3.4 Unemployment estimation

As we explained, to perform the spatial prediction, we created a regular grid of $1km^2$ in the domain. A projection from the mesh to the grid was performed and the resultant map of the posterior mean of the average number of unemployed people per dwelling at location s and quarter t, $\lambda(s,t)$, is shown in figure 4.3. We can see that the average number of unemployed people per dwelling is higher in the Porto, Península de Setúbal and Alentejo regions. We can also distinguish a slight decrease of this indicator over time.

We generate 1000 samples from an approximated posterior of the fitted model, using the INLA function *inla.posterior.sample*, to estimate the target quantities, $E(N(A,t)|\mathbf{x})$, through Monte Carlo sampling. The logarithmic transformation of these quantities are given in figure 4.4. As we might expect, the highest values are in Área Metropolitana de Lisboa and Área Metropolitana do Porto where the population size is higher.

The aggregation of these estimates by NUTS III regions (NUTS-2013) is shown in figure 4.5. Here, we can see that there was a decreasing tendency during the study period.

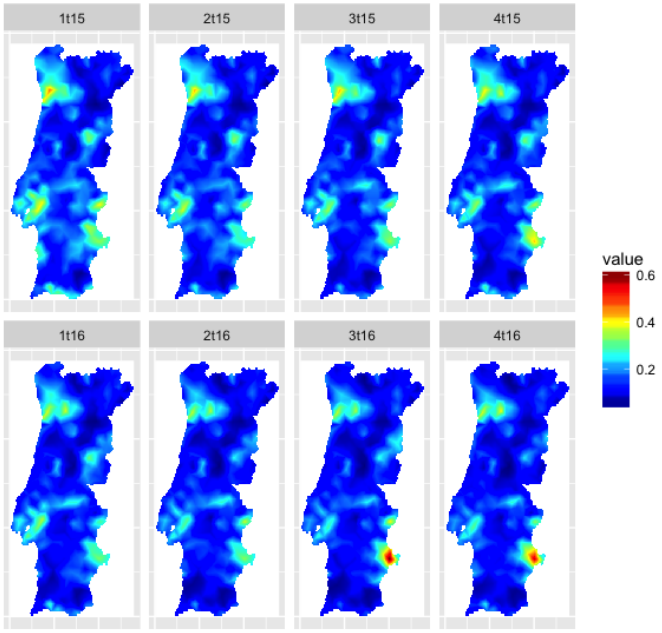


Figure 4.3: Posterior mean of the average number of unemployed people per dwelling by grid cell

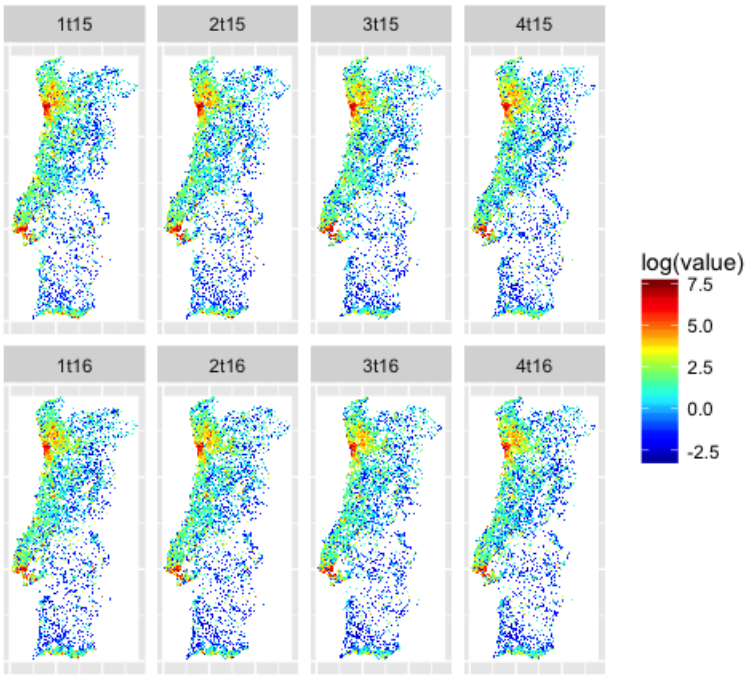


Figure 4.4: Logarithmic transformation of the posterior predictive mean of total unemployed by grid cell

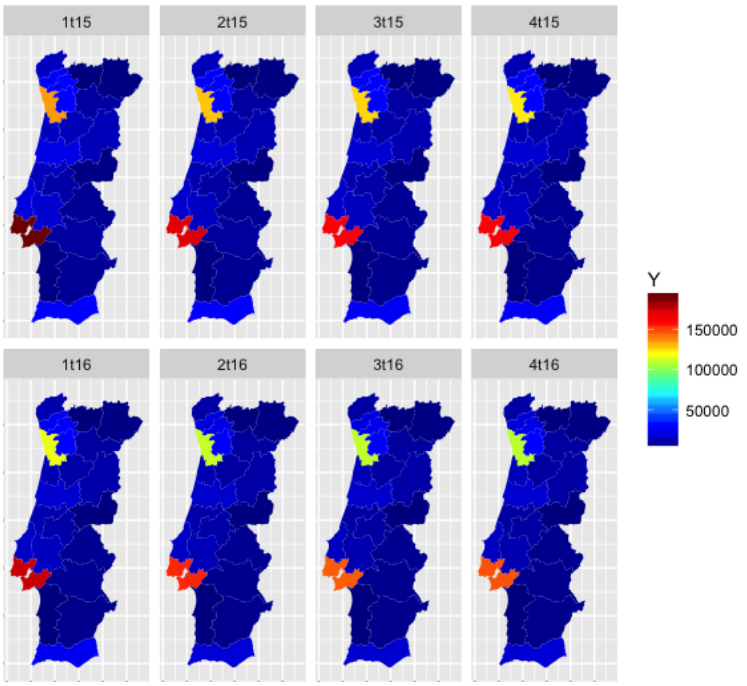


Figure 4.5: Posterior predictive mean of the total unemployed by NUTS III regions

4.4 Sensitivity analysis about the covariates effects

As described in section 3, the covariates considered in the modelling process were the mean age, the median of education level and the proportion of unemployed people registered in the employment centers. These covariates were predicted for the mesh nodes using a Kernel method, with a smoothing bandwidth k equal to 20. We repeated the study with $k = 5$ and the comparison is shown in figure 4.6. The model estimates seem to be sensitive to the kernel parameter used in the covariates prediction. For the majority of regions, the credible intervals are smaller for $k = 5$ (less smooth). Moreover, for $k = 20$ we see over-estimation comparing to the direct method, while for $k = 5$ that relation is not so evident, most noticeably in the Área Metropolitana de Lisboa and Área Metropolitana do Porto regions. In those regions, the population size is high, thus the direct method performs well (as we saw in the previous chapter). Thus, we believe that the choice $k = 5$ would be more appropriate in this study. We also made an analysis of the model without covariates (see credible intervals in figure 4.7). In this case, we see an under-estimation for most regions compared to the direct method. This conclusion reveals the importance of the covariates.

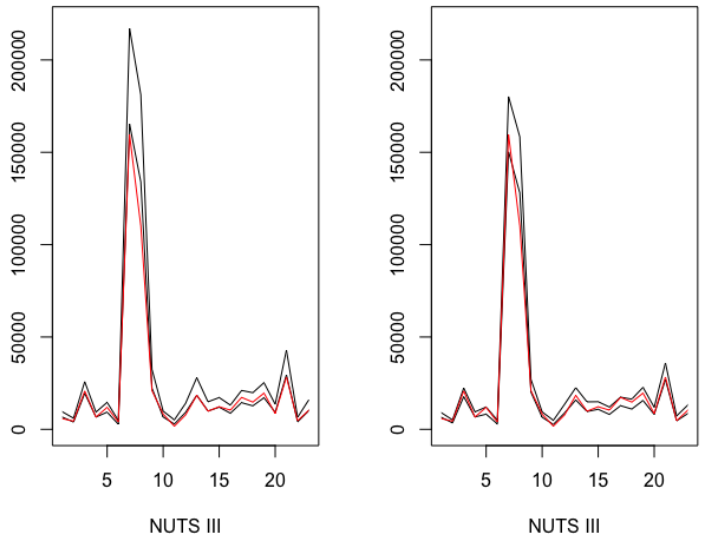


Figure 4.6: 95% Credible intervals for the total unemployed estimates and the direct estimates by NUTS III regions, using $k = 20$ (left) and $k = 5$ (right)

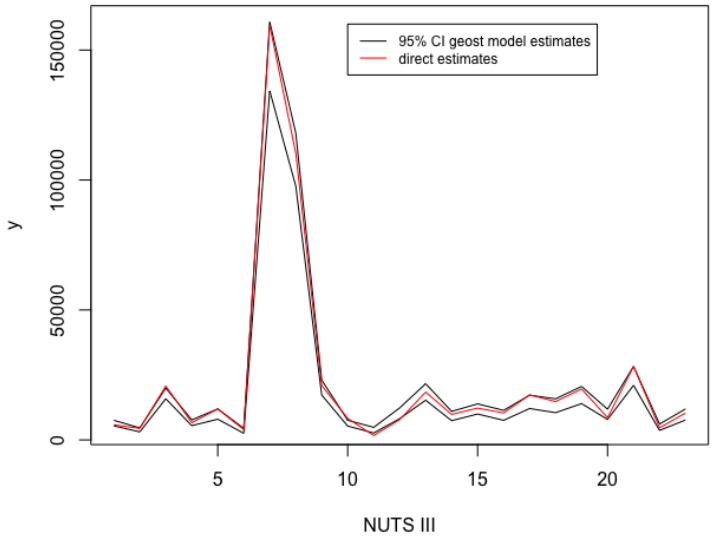


Figure 4.7: 95% Credible intervals for the total unemployed estimates (using the model without covariates) and the direct estimates by NUTS III regions

4.5 Comparison between the results of the geostatistical data model and the traditional small area models

Here we conduct a comparison between the proposed geostatistical data model and the traditional small area models described in the first chapter, specifically the FH model and the FH-CAR model using a Bayesian approach. We applied all models using data from the 4th quarter of 2016 (without temporal extension) to perform the comparison.

Figure 4.8 shows the estimates obtained by the direct method, the FH model, the FH-CAR model and the geostatistical model, along with their respective coefficients of variation, for the NUTS III regions (NUTS-2013). Both the models proposed in the literature, and the geostatistical model, presented lower CVs in comparison with the direct method. As we can see, the model that performed the best in terms of variability was the geostatistical model.

We also can see that the FH models disagree with the direct method in the Área Metropolitana do Porto region (see figure 4.8. This result, in one of the most densely populated regions, may indicate some fragility. Moreover, the coefficients of variation obtained by these models for the Beira Baixa region are quite high (figure 4.10).

The FH models considered here were applied to model the totals of unemployed people in each NUTS III. However, since the regions have different population sizes, it is perhaps more reasonable to model the proportions of unemployed people instead. Indeed, if we do this during the modelling process of FH, we obtain lower

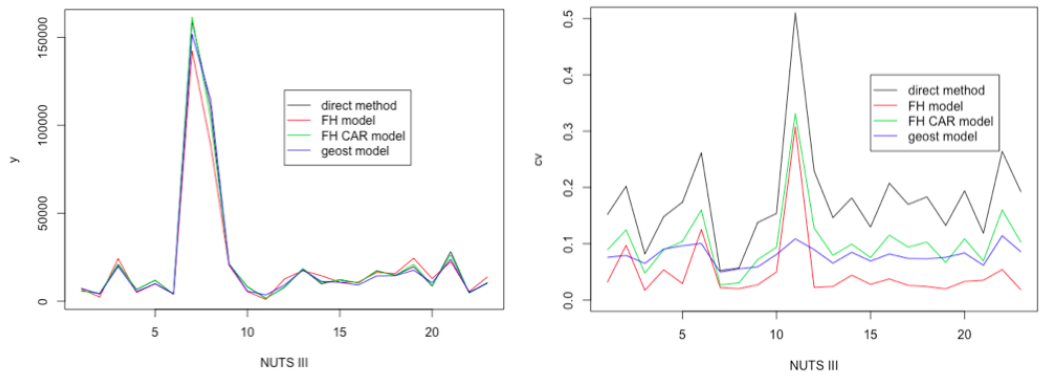


Figure 4.8: Estimates of the total unemployed by NUTS III for the 4th quarter of 2016 and the respective coefficients of variation, for the four models

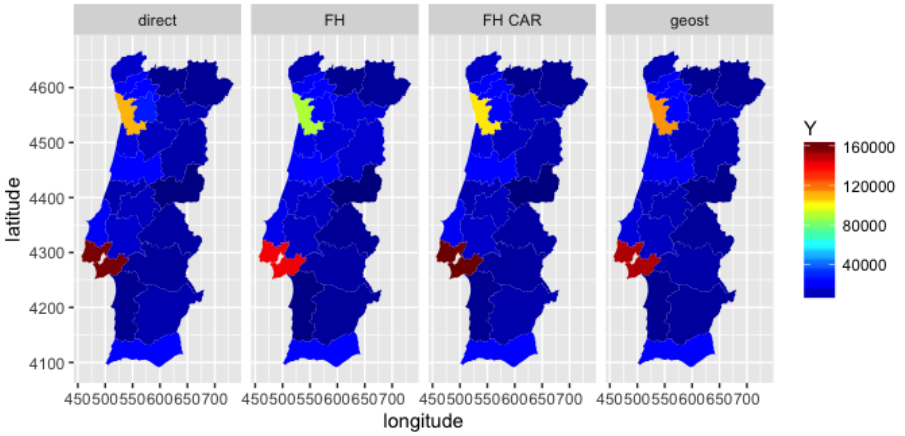


Figure 4.9: Estimates of the total unemployed by NUTS III for the 4th quarter of 2016

coefficients of variation, as shown in figure 4.11. Moreover, the estimates are now closer to the direct ones (figure 4.12). Although modelling proportions using FH models is commonplace in the literature of SAE, the assumption of the normal distribution for proportions is questionable. In the geostatistical model proposed here we modelled the mean of total unemployed in each residential building through a linear predictor using an offset term given by the logarithmic of the number of people in each residential building and some covariates (see section 4.3). Since in this case the data are assumed to be Poisson and the link function used is the log, it is equivalent to model proportions. We think that this approach is clearly more suitable for this problem.

As we expected, the FH model presents lower CVs in comparison with the FH-CAR model and the geostatistical model (figure 4.13). This result can be explained by the difference in the number of parameters in these models. Moreover, the FH models use the direct estimates as the data in the modelling process and assume that the variances are fixed. The geostatistical model uses the data from the sample of the Labour Force Survey, and do not assume the variances as fixed. In any case,

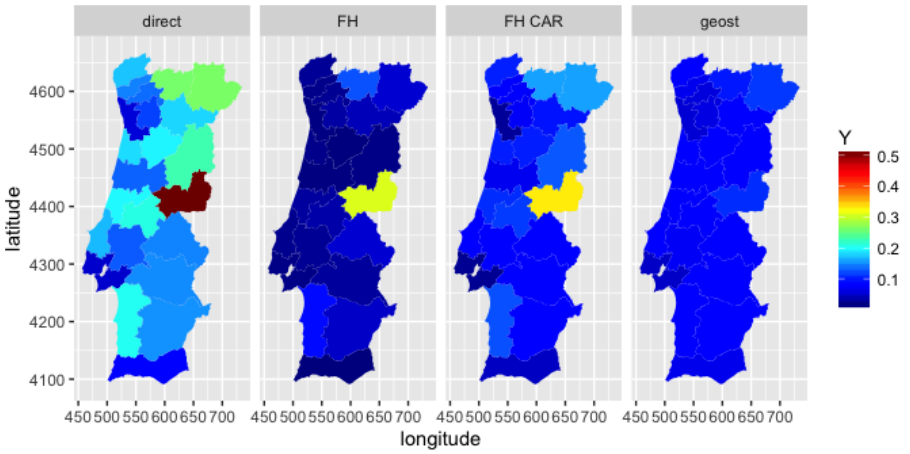


Figure 4.10: Coefficients of variation

the CVs of the estimates obtained by the FH models, as well as the CVs of the estimates obtained by the geostatistical model do not achieve 20% in any NUTS III region.

Since the estimates and the CVs of these three models are close, we think that the geostatistical model proposed here brings many advantages to this problem, as we explained in the previous section.

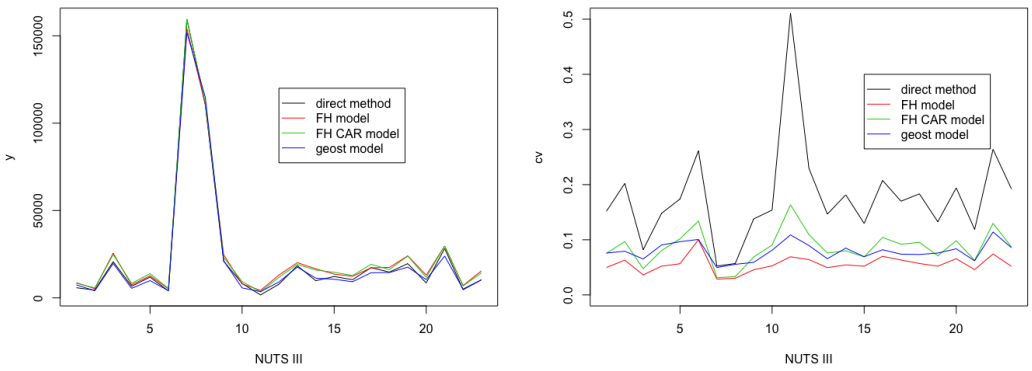


Figure 4.11: Estimates of the total unemployed by NUTS III for the 4th quarter of 2016 and the respective coefficients of variation, for the four models

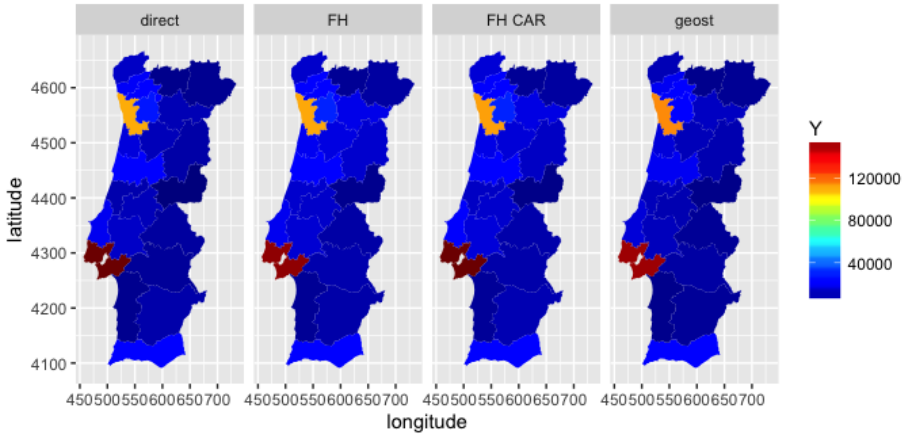


Figure 4.12: Estimates of the total unemployed by NUTS III for the 4th quarter of 2016

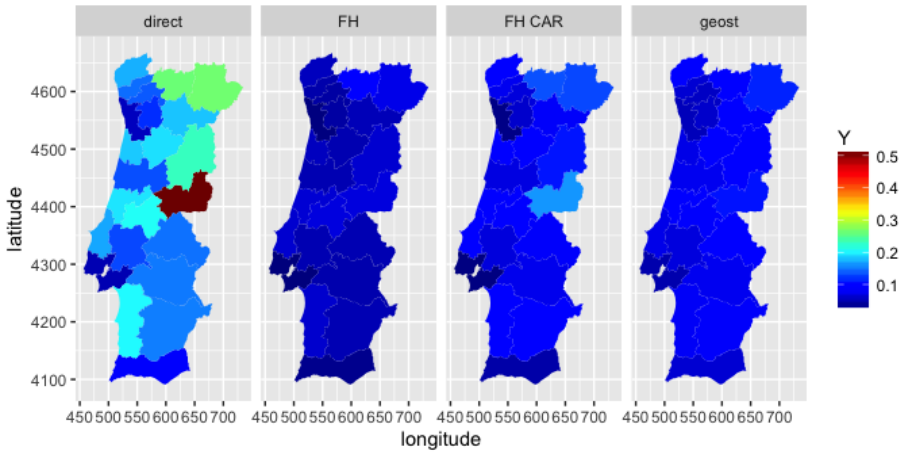


Figure 4.13: Coefficients of variation

4.6 Comparison between the results of the three approaches presented in chapters 2, 3 and 4

For a comparison between the three proposed approaches, we considered the respective spatial versions (without temporal extensions), using only data from the 4th quarter of 2016. The new version of the geostatistical data model uses the same mesh as was used in the spatial point processes model to be comparable. We also considered the same set of covariates in the three approaches.

Figure 4.14 shows the estimates of the total number of unemployed people by NUTS III regions (NUTS-2013) for the 4th quarter of 2016, using the direct method and the three approaches, and the respective coefficients of variation. Although significant differences are not visible in the estimates produced by each of the models, the coefficients of variation are distinguishable. The three proposed models provided estimates that are more accurate than the direct method. Usually the Portuguese

National Statistical Office requires CVs lower than 20% for the estimates to be published as official figures. The three methods proposed respect and adhere to this requirement. The areal model is the one with the lowest CVs. However, it is important to note that the LGCP model and the geostatistical data model ('geost' in the graph) provide much more information than the areal model. Intuitively, we can say that whereas the direct method and the areal model only tell us how many unemployed there are, the LGCP model and the geostatistical model tell us how many and where are they. Therefore, we expect more variability in these models than in the areal model. Moreover, since the point-referenced data model does not require the modelling of the points, we expect less variability in comparison with the spatial point processes model.

Figure 4.15 permits a better analysis of the results in space. Here, we can see that the point referenced data model is the closest model to the direct method in the Área Metropolitana de Lisboa and Área Metropolitana do Porto regions. These are the regions with the highest population density, and we know that the sample size is large enough to provide accurate estimates in these regions using the direct method (see regions 7 and 8 in figure 4.14). We also know that the direct estimator is unbiased (see chapter 1). Thus, the correlation with the direct estimates in these regions is a favourable result.

The regions with the highest CVs using the direct method are Beira Baixa, Terras de Tras-os-Montes and Alto Tamega (see figure 4.16). Although the areal model is the one with the lowest CVs on average, its maximum value (in Alto Tâmega) is close to the highest value using the geostatistical model.

In addition to these results, it is important to note that the LGCP model and the geostatistical model bring many advantages in comparison with the direct method and areal data models. We list some of these advantages: the possibility of providing estimates for all municipalities or in even more detailed geographical regions, the coherence between different geographical levels, and the provision of information about the number of unemployed people per residential building, while taking into account specific information about the families.

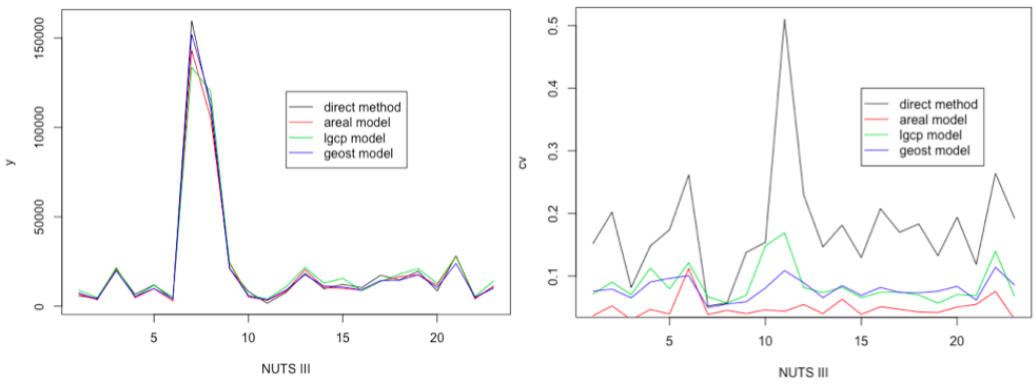


Figure 4.14: Estimates of the total unemployed by NUTS III for the 4th quarter of 2016, and the respective coefficients of variation

Table 4.2 provides a summary of the characteristics of the proposed models, as

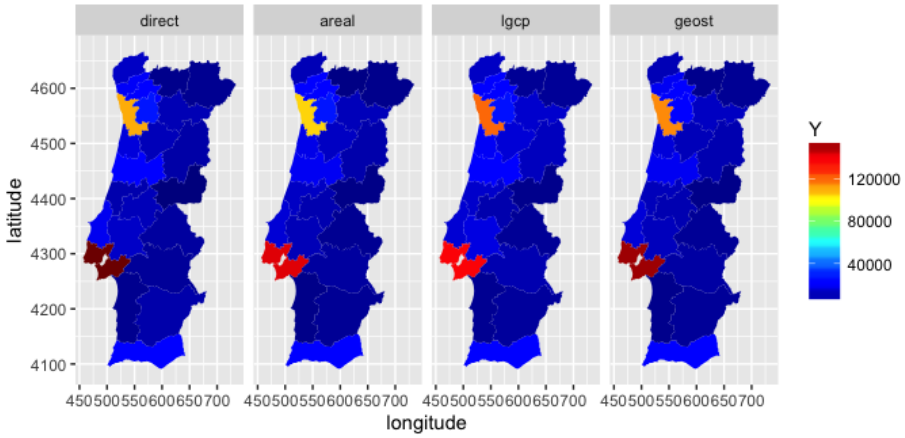


Figure 4.15: Estimates of the total unemployed by NUTS III for the 4th quarter of 2016

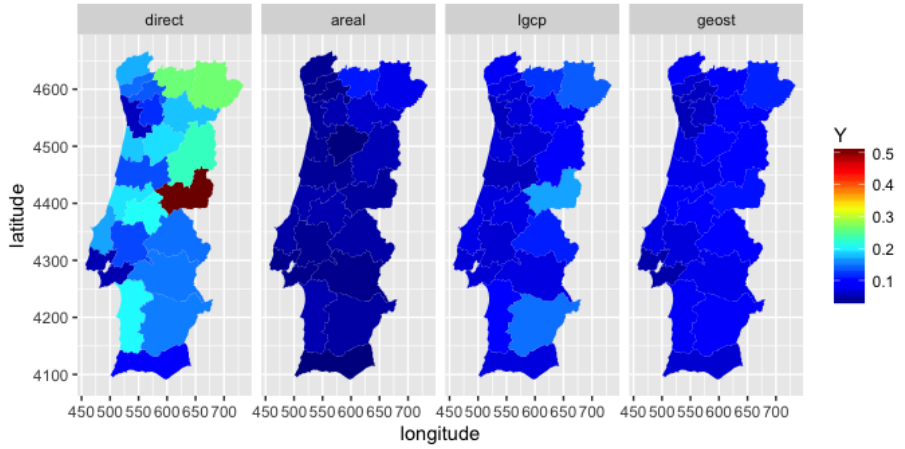


Figure 4.16: Comparison between the coefficients of variation of the estimates using the four models

well as the traditional SAE models.

One of the most significant differences between the areal and point-referenced models is the extent of the spatial effects. In an areal model, the spatial dependence is taken into account through the structure of a CAR model, whereas a point-referenced model uses a Gaussian field, which is much closer to reality in the majority of applications.

We can also note that the traditional methods in SAE assume a normal distribution for the variable of interest. Sometimes, this assumption is undesireable and generalized linear models must be considered. Moreover, they use the direct estimates as observations and consider that the variances are fixed. However, on the other hand, the proposed models use the observations of the LFS sample as data and consider a Poisson distribution for the counts.

Another difference between the areal and point-referenced models is the level of

the covariates. Areal level models can only use covariates at area level, whereas point-referenced models can use more specific information at point level (for distinct residential buildings for example). Naturally, with more detailed data and information, more accurate and precise estimates can be produced.

One of the most important issues in the SAE is the sampling design. In most cases, we try to do inference for the population using information from the sample. Thus, the sample must be representative of the whole population. However, it is common to find units in the sample with different selection probabilities. Therefore, some units must be more weighted than others in the estimation process. In the Portuguese LFS the probabilities are homogeneuous inside each grid cell. Since the traditional SAE methods use the direct estimates as input, they take into account these probabilities at grid cell level. The areal models proposed here in chapter 2 only take into account the selection probabilities at area level, assuming the same weights for the observations at different grid cells in the same NUTS III region. That assumption may lead naturally to biased estimates. However, the proposed models in chapters 3 and 4 solve this problem. The predictive intensities are multiplied by the inverse of selection probabilities at $1km^2$ grid cell to obtain the estimates for the population. Thus, the level of the used sampling design information is the same as for the traditional SAE methods.

The point-referenced models can provide much more detailed information, specifically at a spatial resolution of the $1km^2$. Consequently, these methods can produce estimates for all municipalities in mainland Portugal. Conversely, since the traditional SAE methods use direct estimates as input, and given that such estimates do not exist for municipalities where there are no observations in the sample, these methods can not produce estimates for all municipalities with satisfactory precision.

In addition to the detailed geographical level of the output, the point referenced methods provide a probability structure containing much more information than the areal models. Whereas the areal models can produce information about the number of unemployed people in areal units, the point-referenced models can provide not only this, but also more specifically, information about where those people are.

In general, the National Statistical Institutes require consistency between hierarchical geographical levels. This means that the estimates for the municipalities inside a given NUTS III region must collectively total the estimate obtained for that entire region. Since the estimates for a given region using the point referenced methods are calculated through the integral of the intensity process, that property is guaranteed. The same is not true for the areal models however.

	SAE methods (HB)		
Model	FH	FH-CAR	Battese *et al.* (1988)
Spatial effects	No	Yes	No
Level of spatial eff.	-	CAR	-
Level of input	Area	Area	Individual
Input	$\hat{\theta_i}, x_i$	$\hat{\theta_i}, x_i$	$y_{ij}, x_{ij}, \bar{X}_i$
Assumptions	Normality of $\hat{\theta_i}$	Normality of $\hat{\theta_i}$	Normality of y_{ij}
Coordinates	-	-	-
Level of covariates	Area	Area	Individual
Level samp. design	cell $1km^2$	cell $1km^2$	cell $1km^2$
Level of output	Area	Area	Area
Spatial resolution	Area	Area	Area
Output	posterior marginals of θ_i	posterior marginals of θ_i	posterior marginals of θ_i
Municip. level	No	No	No
Consistency	No	No	No
Probability structure	number of unemployed in areal units	number of unemployed in areal units	number of unemployed in areal units
	Proposed methods (HB)		
Model	Areal model	LGCP model	Geostatistical model
Spatial effects	Yes	Yes	Yes
Level of spatial eff.	CAR	Gaussian field	Gaussian field
Level of input	Area	Residential building	Residential building
Input	y_i, x_i	$s, y(s), x(s)$	$s, y(s), x(s), S$
Assumptions	Poisson distribution for y_i	LGCP for points, poisson for marks	Poisson distribution for $y(s)$
Coordinates	-	Random	Fixed
Level of covariates	Area	Residential building and area	Residential building and area
Level samp. design	Area	cell $1km^2$	cell $1km^2$
Level of output	Area	continuous space in the domain	continuous space in the domain
Spatial resolution	Area	cell $1km^2$	cell $1km^2$
Output	posterior marginals of λ_i	posterior marginals of $\lambda_1(s)$ and $\lambda_2(s)$	posterior marginals of $\lambda(s)$
Municip. level	No	Yes	Yes
Consistency	No	Yes	Yes
Probability structure	number of unemployed in areal units	Random configuration of sampling units and their marks in space + corresponding counting processes	Random configuration of marks in space + corresponding counting processes

Table 4.2: Comparison between HB SAE methods and proposed models

4.7 Discussion

In this study we looked at unemployment data from a new perspective. In most National Statistical Institutes, the unemployment estimation is made using a direct method. However, although some of these institutes are starting to use small area estimation techniques to produce accurate estimates using areal models, these models do not take into account specific information about the households, and the geographical information is not sufficiently detailed.

In the previous chapter, we proposed to look at unemployment through a marked spatial point process (Pereira et al, 2017), where the points are the locations of the residential buildings and the marks attached are the total unemployed in each point.

However, during this time the locations of all residential buildings in the national territory became available, meaning there is now no need to model its intensity, producing extra variability. Here, we proposed to look at unemployment data as geostatistical data, assuming that all locations of the residential buildings are known. Moreover, we considered a spatio-temporal extension, using data from the 4th quarter of 2014 to the 4th quarter of 2016.

This methodology not only provides, with a high degree of accuracy, the unemployment estimates for every quarter in the study, but also for every area (municipalities, NUTS, etc.) using the same model in a consistent way.

We also concluded that the inclusion of covariates is very important. Moreover, the model is sensitive to the smoothing parameter in the kernel smoothing used to perform the spatial prediction of the covariates included in the model. Therefore, we suggest that in these cases, a sensitive analysis must be conducted.

A comparison with the direct method, the areal model, the marked LGCP model, and the traditional SAE methods showed that the geostatistical model is one of

the models with best performance in terms of bias and variance, and has many advantages when employed for estimation in small areas.

For future investigation, we think it would be interesting to do an elicitation of priors for the hyperparameters and compare these results with those obtained using PC priors.

Bibliography

[Adler *et al.* (2013)] Adler, D., Kneib, T., Lang, S., Umlauf, N., Zeileis, A. (2013). BayesXsrc: R Package Distribution of the BayesX C++ Sources. R package version 2.1-2. https://cran.r-project.org/web/packages/BayesXsrc/

[Baddeley *et al.* (2016)] Baddeley, A., Rubak, E., Turner, R. (2016) *Spatial Point Patterns - Methodology and Applications with R.* Chapman and Hall/CRC.

[Banerjee *et al.* (2004)] Banerjee, S., Carlin, B. P., Gelfand, A. E. (2004) *Hierarchical modelling and Analysis for Spatial Data.* Chapman and Hall/CRC.

[Banerjee *et al.* (2015)] Banerjee, S., Carlin, B. P., Gelfand, A. E. (2015) *Hierarchical modelling and Analysis for Spatial Data*, 2nd edition. Chapman and Hall/CRC.

[Bardenet (2017)] Bardenet, R., Doucet, A. (2017) On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research* **18**, 1-43.

[Besag (1991)] Besag, J., York, J., Mollie, A. (1991) Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math* **43**, 1-59.

[Best (2005)] Best, N., Richardson, S., Thomson, A. (2005) A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research* **14**, 35-59.

[Blangiardo *et al.* (2015)] Blangiardo, M., Cameletti, M. (2015) *Spatial and Spatio-temporal Bayesian Models with R-INLA*. Wiley.

[Bonneu *et al.* (2007)] Bonneu, F. Exploring and modelling fire department emergencies with a spatio-temporal marked point process (2007). *Case Studies in Business, Industry and Government Statistics*, **1**, 139-152.

[Brooks and Gelman (1997)] Brooks, S. P., and A. Gelman (1997). Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455.

[Choundry *et al.* (1989)] Choundry, G. H., Rao, J.N.K. Small area estimation using models that combine time series and cross sectional data (1989). *Journal of Statistics Canada Symposium on Analysis of Data in Time* 67-74.

[Cressie (1991)] Cressie, N. (1991) *Statistics for spatial data*, Wiley Series.

[Da-Silva (2016)] Da-Silva, C.Q., and Migon,H.S.(2016) Hierarchical Dynamic Beta model. *REVSTAT*, **14**, 49-73.

[Datta *et al* (1999)] G.S. Datta, P. Lahiri, T. Maiti, K.L. Lu (1999) Hierarchical Bayes estimation of unemployment rates for the US states. *Journal of the American Statistical Association*, **94**, 1074-1082.

[Datta *et al.* (1991)] Datta, G. S. and Ghosh, M. Bayesian prediction in linear models: application to small area estimation (1991). *Ann. Statist.* **19**, 1748-1770.

[Deville and Sarndal (1992)] Deville, J. C., Sarndal, C. E. (1992) Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, **87**, 376-382.

[Diggle (2003)] Diggle, P. (2003) *Statistical Analysis of Spatial Point Patterns.* Arnold.

[Diggle *et al.* (1998)] Diggle, P. J., Tawn, J. A., Moyeed, R. A. Model-based geostatistics (1998). *Appl. Statist.* **47**, 299-350.

[Fay *et al.* (1979)] Fay, R.E, Herriot,R.A.(1979) Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269-277.

[Ferrari *et al.* (2004)] Ferrari, S., Cribari-Neto, F. (2004) Beta Regression for modelling Rates and Proportions. *Journal of Applied Statistics*, **31**, 799-815.

[Fuglstad *et al.* (2017)] Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2017) Constructing Priors that Penalize the Complexity of Gaussian Random Fields. arXiv:1503.00256

[Gelfand *et al.* (2004)] Gelfand, A. E., Schmidt, A.M., Banerjee, S. and Sirmans, C.F. (2004) Nonstationary multivariate process modelling through spatially varying coreginalization. *Test*, **13**, 263-312.

[Gelman *et al.* (2014)] Gelman, A., Hwang, J., and Vehtari, A. (2014) Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, **24**, 997-1016.

[Gelman et al (2004)] Gelman, A.; Carlin, J.; Stern, H. and Rubin, D. (2004) *Bayesian Data Analysis, Second Edition*, Chapman and Hall.

[Gneiting *et al.* (2007)] Gneiting, T., Raftery, A.E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal American Statistical Associaton.* **102**, 359-378.

[Horvitz and Thompson (1952)] Horvitz, D. G., Thompson, D. J. (1952) A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, **47**, 663-685.

[Illian *et al.* (2012a)] Illian, J. B., Sørbye, S. H., Rue, H. and Hendrichsen, D. K. (2012) Using INLA to fit a complex point process model with temporally varying effects – a case study. *Journal of Environmental Statistics*, **3**.

[Illian *et al.* (2012b)] Illian, J. B., Sorbye, S. H., Rue, H., (2012) A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *The Annals of Applied Statistics*, **6**, 1499-1530.

[Illian *et al* (2008)] Illian, J., Penttinen, A., Stoyan, H., Stoyan, D. (2008) *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley.

[Illian and Rue (2010)] Illian, J.B., Rue, H. (2010) A toolbox for fitting complex spatial point process models using integrated Laplace transformation (INLA). *Technical Report, Trondheim University*.

[INE (2015)] INE (2015) NUTS III - As novas unidades territoriais para fins estatísticos. *Technical Report*.

[Jedrzejcczak and Kubacki (2017)] Jedrzejcczak, A., Kubacki, J. (2017) ESTIMATION OF SMALL AREA CHARACTERISTICS USING MULTIVARIATE RAO-YU MODEL. *Statistics in Transition*, **18**, 725-742.

[Jeffreys (1946)] Jeffreys, H. (1946) An invariant form for the prior probability in estimation problems . *Proceedings of the Royal Society A*, **186**, 453-461.

[Jiang *et al.* (2006a)] Jiang, J. and Lahiri, P. Estimation of finite population domain means: A model-assisted empirical best prediction approach (2006a). *J. Amer. Statist. Assoc.* **101**, 301-311.

[Jiang *et al.* (2006b)] Jiang, J. and Lahiri, P. Mixed model prediction and small area estimation (2006b). *Test 15*, 1-96.

[Knorr-Held (2000)] Knorr-Held, L. (2000) Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, **19**, 2555-2567.

[Krainski *et al.* (2017)] Krainski, E., Lindgren, F., Simpson, D., Rue, H. (2017) The R-INLA tutorial on SPDE models. http://www.math.ntnu.no/inla/r-inla.org/tutorials/spde/spde-tutorial.pdf

[Lawson (2009)] Lawson, A. (2009) *Bayesian disease mapping*, Chapman & Hall/CRC.

[Lindgren *et al.* (2011)] Lindgren, F., Rue, H., Lindstrom, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the SPDE approach (with discussion). *Journal of Royal Statistical Society Series B*, **73**, 423-498.

[Lopez-Vizcaino *et al.* (2015)] Lopez-Vizcaino, E., Lombardia, M. J., Morales, D. (2015) Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *Journal of Royal Statistical Society Series A*, **178**, 535-565.

[Lopez-Vizcaíno *et al* (2013)] Lopez-Vizcaíno, E., Lombardía, M. J., Morales, D. (2013) Multinomial-based small area estimation of labour force indicators. *Statistical Modelling*, **13**, 153-178.

[Lunn *et al.* (2000)] Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D. J. (2000) WinBUGS - A Bayesian modelling framework: concepts, structure and extensibility. *Statistics and Computing*, **10**, 325–337.

[Marhuenda *et al.* (2013)] Marhuenda, Y., Molina, I., Morales, D. (2013) Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, **58**, 308-325.

[Martino *et al.* (2010)] Martino S. and Rue H. (2010) Case Studies in Bayesian Computation using INLA. *Complex data modelling and computationally intensive statistical methods* (R-code).

[Martins *et al.* (2013)] Martins, T., G., Simpson, D., Lindgren, F., Rue, H. (2013) Bayesian computing with INLA: New features. *Computational Statistics and Data Analysis*, **67**, 68-83.

[Molina *et al.* (2007)] Molina,I., Saei, A. and Lombardia, M.J. (2007) Small area estimates of labour force participation under a multinomial logit mixed model. *J.R. Statist. Soc. A*, **170**, 975-1000.

[Moller and Waagepetersen (2003)] Moller, J., Waagepetersen, R. P. (2003) *Statistical Inference and Simulation for Spatial Point Processes.* Wiley.

[Nadaraya (1989)] Nadaraya, E. (1989) *Nonparametric estimation of probability densities and regression curves.* Vol. 20 of Mathematics and its Applications (Soviet Series), Kluwer Academic Publishers Group, Dordrecht. Translated from the Russian by Samuel Kotz.

[Nadaraya (1964)] Nadaraya, E. (1964) On estimating regression. *Theory of Probability and its Applications*, **9**, 157–159.

[Neyman and Scott (1958)] Neyman, J. and Scott, E. L. (1958) Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society B.*, **20**, 1-29.

[O'Hagan *et al* (2006)] O'Hagan, A., Buck, C., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., Oakley, J., Rakow, T. (2006) *Uncertain Judgements: Eliciting Experts' Probabilities.* Wiley

[Paulino *et al.* (2003)] Paulino, C. D., Turkman, M. A., Murteira, B. (2003) *Estatística Bayesiana* (1st edition), Fundação Calouste Gulbenkian.

[Pereira (2014)] Pereira, P. (2014) *Métodos Probabilísticos e Estatísticos na Gestão de Fogos Florestais* . PhD thesis. http://gi3ceaul.fc.ul.pt/Report3/Tese_ Paula_ Pereira.pdf

[Pereira *et al.* (2016)] Pereira, S., Turkman, F., Correia, L. (2016) Spatio-temporal analysis of regional unemployment rates: A comparison of model based approaches. To appear in *Revstat*. https://arxiv.org/abs/1704.05767

[Pereira *et al.* (2017)] Pereira, S., Turkman, F., Correia, L., Rue, H. (2017) Unemployment estimation: Spatial point referenced methods and models. https://arxiv.org/pdf/1706.08320.pdf

[Pfeffermann *et al.* (2002)] Pfeffermann, D. (2002) Small area estimation: new developments and directions, *Int. Statist. Rev.*, **70**, 125-143.

[Plummer (2003)] Plummer, M. (2003) JAGS: A program for analysis of Bayesian graphical models using GIBBS sampling.

[Pratesi and Salvati (2008)] Pratesi, M., Salvati, N. (2008) Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Stat. Methods Appl.*, **17**, 113-141.

[Rao (1994)] Rao, J. N. K., Yu, M. (1994) Small area estimation by combining time series and cross sectional data *Canadian Journal of Statistics*, **22**, 511-528.

[Rao (2003)] Rao, J.N.K. (2003) *Small Area Estimation*. New York: Wiley.

[Rao and Molina (2015)] Rao, J.N.K., Molina, I. (2015) *Small Area Estimation*, Second edition. New York: Wiley.

[Roos (2015)] Roos, M., Martins, T. G., Held, L., and Rue, H. (2015) Sensitivity Analysis for Bayesian Hierarchical Models. *Bayesian Analysis*, **2**, 321-349.

[Rue *et al.* (2009)] Rue, H., Martino, S., Chopin, N. (2009) Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion). *Journal of the Royal Statistical Society Series B*, **71**, 319-392.

[Rue *et al.* (2017)] Rue, H., Riebler, A., Sorbye, S. H., Illian, J. B., Simpson, D. P., Lindgren, F. K. (2017) Bayesian computing with INLA: A review. *Annual Reviews of Statistics and Its Applications*, **4**, 395-421.

[Schoenberg (2004)] Schoenberg, F. P. (2004) Testing separability in spatial-temporal marked point processes. *Biometrics*, **60**, 471-481.

[Simpson *et al.* (2017)] Simpson, D. P., Rue, H. , Riebler, A. , Martins, T. G. , and Sørbye, S. H. (2017) Penalising model component complexity: A principled, practical approach to constructing priors (with discussion). *Statistical Science*, **32**, 1-28.

[Simpson et al.(2016)] Simpson, D. , Illian, J. , Lindgren, F. , Sørbye, S. , and Rue, H. . (2016) Going off grid: Computational efficient inference for log-Gaussian Cox processes. *Biometrika*, **103** ,1-22, 2016. (doi: 10.1093/biomet/asv064).

[Singh (2005)] Singh, B., Shukla, G., Kundu, D. Spatio-temporal models in small area estimation (2005), *Survey Methodology*, **31**, 183-195.

[Spieghalter *et al.* (2002)] Spiegelhalter, D. J., Best, N.G., Carlin, B.R., van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of Royal Statistical Society Series B*, **64**, 583-639.

[Stan Development team (2014)] Stan Development team (2014) Stan: A C++ Library for Probability and Sampling, Version 2.5.0. http://mc-stan.org

[Thomas et al (2006)] Thomas, A., O'Hara, B., Ligges, U., Sturtz, S. (2006) Making BUGS Open. *R News*, **6**, 12–17.

[Tobler (1970)] Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, **46**, 234–240.

[Turkman *et al.* (2015)] Turkman, M. A., Paulino, C. D. (2015) *Estatística Bayesiana Computacional - uma introdução.* Sociedade Portuguesa de Estatística.

[van der Vaart (2000)] van der Vaart, A., W. (2000) Asymptotic Statistics. *Cambridge University Press.*

[Watanabe, S. (2010)] Watanabe, S. (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* **11**, 3571-3594.

[Watson (1964)] Watson, G. S. (1964) Smooth regression analysis. *Sankhya Ser. A*, **26**, 359-372.

[Wolter (2007)] Wolter, K. M. (2007) *Generalized Variance Functions.* Springer

[You *et al.* (2006)] You,Y., Chapman, B. (2006) Small area estimation using area level models and estimated sampling variances *Survey methodology*, **32**, 97-103

[You *et al.* (2011)] You, Y. and Zhou, Q.M. (2011) Hierarchical Bayes small area estimation under a spatial model with application to health survey data. *Survey methodology*, **37**, 25-237.

[You and Gambino (2003)] You, Y., Rao, J., and Gambino, J. (2003) Model-Based Unemployment Rate Estimation for the Canadian Labour Force Survey: A Hierarchical Bayes Approach. *Survey Methodology*, **29**, 25-32.

[Yuan *et al.* (2017)] Y. Yuan, F. E. Bachl, F. Lindgren, D. L. Brochers, J. B. Illian, S. T. Buckland, H. Rue, T. Gerrodette. (2017) Point process models for spatio-temporal distance sampling data. *Annals of Applied Statistics.*

Appendix A

NUTS III - version 2013

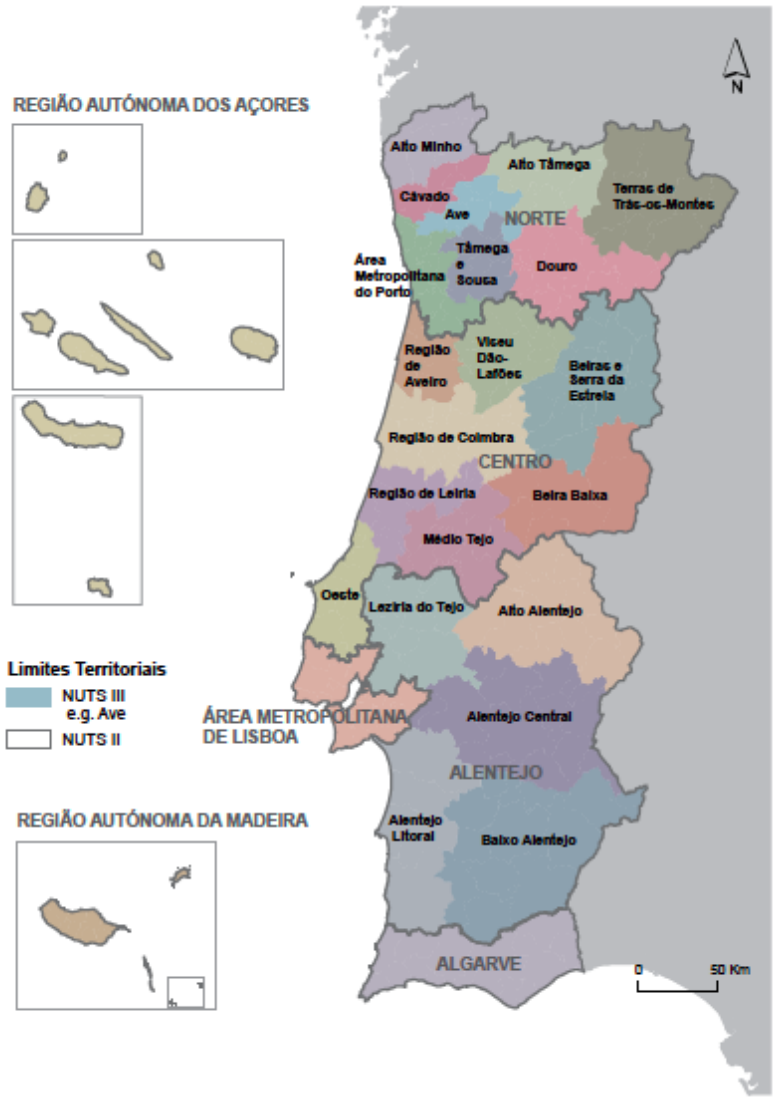


Figure A.1: NUTS III - version 2013 (figure taken from INE, 2015)

Appendix B

R-codes and programs

B.1 Small area estimation methods

B.1.1 Data preparation

```
#data from the 4th quarter of 2014

trim<-"4t14"

p<-paste(getwd(),"/dados/IE_pais_",trim,".csv",sep="")
p

base_dados<-read.table(paste(p),head=T,sep=",",dec=".",fileEncoding="
    ↪ latin1")
base_dados[1:3,]

pesos<-read.table(paste(getwd(),"/dados/pesos",trim,".txt",sep=""),
    ↪ head=TRUE,fileEncoding="latin1")


dados_1t<-merge(base_dados,pesos,by="IDENTIF_FIXA_COMPLETA")
dim(dados_1t)

#remove the islands

dados_1t<-dados_1t[which(dados_1t$NUTS.3.NG!=200&dados_1t$NUTS.3.NG
    ↪ !=300),]

#mean age by NUTS III

idade<-tapply(dados_1t$IDADE_ANOS,droplevels(dados_1t$NUTS.3.NG),mean
    ↪ )
idade<-c(t(idade))
```

```
#shapefile with municipalities in mainland Portugal

concelhos<-readRDS("concelhos.rds")



#population estimates by municipality for 2014

pop_concelho<-read.table(paste(getwd(),"/dados/pop_concelho2014.txt",
    ↪ sep=""))


names(pop_concelho)<-c("codigo","pop")


library(stringr)
pop_concelho$dico<-str_sub(pop_concelho$codigo, start= -4)


#data from the IEFP by municipality for the 4th quarter of 2014

iefp_fev<-read.table(paste(getwd(),"/dados/iefp_nov14.txt",sep=""),
    ↪ colClasses=c("character","numeric"))


names(iefp_fev)<-c("dico","iefp")

iefp_fev2<-merge(pop_concelho,iefp_fev,by="dico")

iefp_fev2$prop_iefp<-iefp_fev2$iefp/iefp_fev2$pop


#shapefile with NUTS III regions in mainland Portugal

nuts3<-readRDS("nuts3.rds")
summary(nuts3)


#municipalities over NUTS III

library(rgeos)
trueCentroids = gCentroid(concelhos,byid=TRUE)

centroides<-trueCentroids
```

```
coord_centroides<-coordinates(centroides)

coords_data <- SpatialPoints(coord_centroides)
proj4string(coords_data)<-proj4string(nuts3)


match_coords_asl <- over(coords_data,nuts3)


#data from IEFP by NUTS III

match_coords_asl$xmean_s<-c(iefp_fev2$iefp)
match_coords_asl$xmean_s2<-c(iefp_fev2$pop)

iefp_nuts3<-aggregate( match_coords_asl$xmean_s ~
   ↪ match_coords_asl$NUTS3_02, FUN = sum )
pop_nuts3<-aggregate( match_coords_asl$xmean_s2 ~
   ↪ match_coords_asl$NUTS3_02, FUN = sum )

names(iefp_nuts3)<-c("NUTS3","iefp")
names(pop_nuts3)<-c("NUTS3","pop")

pop2014<-pop_nuts3$pop

prop_iefp_nuts3<-data.frame(pop_nuts3$NUTS3,iefp_nuts3$iefp/
   ↪ pop_nuts3$pop)

names(prop_iefp_nuts3)<-c("NUTS3","prop_iefp")

iefp4t14<-prop_iefp_nuts3$prop_iefp

iefp_totals<-iefp4t14*pop2014

#standardized data

idade2<-(idade-mean(idade))/sd(idade)
iefp2<-(iefp4t14-mean(iefp4t14))/sd(iefp4t14)
pop2<-(pop2014-mean(pop2014))/sd(pop2014)

dados<-data.frame(diretas,idade2,iefp2,pop2,iefp_totals)


#precision of direct estimates to enter in the FH model

prec0<-1/((1/10000^2)*var_dir)
```

B.1.2 FH

```
#FH model

formula <- diretas/10000 ~ 1 + idade2 + iefp2 + pop2

library(INLA)

fh <-inla(formula,scale=prec0,family="gaussian",data=dados,
          control.family=list(hyper=list(prec = list(initial = 1,
             ↪ fixed=TRUE))),
          control.predictor=list(compute=TRUE), control.compute=list(
             ↪ dic=TRUE,cpo=TRUE,waic=TRUE,config=TRUE),
          control.inla=list(strategy="gaussian"))

#posterior mean

media_fh<-fh$summary.fitted.values$mean*10000

#posterior standard deviation

sd_fh<-fh$summary.fitted.values$sd*10000

#coefficients of variation

cv_fh<-sd_fh/media_fh
cv_fh
```

B.1.3 FH-CAR

```
#adjacency matrix for NUTS III regions

nuts3_2<-nuts3[order(nuts3$NUTS3_02),]

library(spdep)
temp2 <- poly2nb(nuts3_2)

nb2INLA("nuts3_2.graph", temp2)
nuts3_2.adj <- paste(getwd(),"/nuts3_2.graph",sep="")

dados$area<-1:28
```

```
#FH-CAR model

formula <- diretas/10000 ~ 1 + idade2 + iefp2 + pop2 + f(area,model="
   ↪ besag",graph=nuts3_2.adj)

fh_s <-inla(formula,scale=prec0,family="gaussian",data=dados,
         control.family=list(hyper=list(prec = list(initial = 1,
            ↪ fixed=TRUE))),
         control.predictor=list(compute=TRUE), control.compute=list(
            ↪ dic=TRUE,cpo=TRUE,waic=TRUE,config=TRUE),
         control.inla=list(strategy="gaussian"))


#posterior mean

media_fh_s<-fh_s$summary.fitted.values$mean*10000

#posterior standard deviation

sd_fh_s<-fh_s$summary.fitted.values$sd*10000

#coefficients of variation

cv_fh_s<-sd_fh_s/media_fh_s
cv_fh_s
```

B.2 Areal data models

B.2.1 Poisson

```
data_paper[1:3,]

library(INLA)

#Poisson model

formula.ST2_poisson<- y ~ empresas+setor_primario+setor_secundario+
   ↪ PIB_trim+iefp +sa6+sa8+
  f(ID.area,model="besag",graph=portugal.adj) +
  f(ID.year,model="rw1")+f(ID.area.year,model="iid")

model.inla.ST2_poisson <-inla(formula.ST2_poisson,family="poisson",
   ↪ data=data_paper,
```

```
                              offset=log(n),control.predictor=list(
                                 ↪ compute=TRUE), control.compute=list
                                 ↪ (dic=TRUE,cpo=TRUE))


#PIT and CPO values
par(mfrow=c(1,2))
pit_poisson<-model.inla.ST2_poisson$cpo$pit
plot(pit_poisson,xlab="Dominio",ylab="PIT")
hist(pit_poisson,xlab="PIT",ylab="Frequencia",main="")
-mean(log(model.inla.ST2_poisson$cpo$cpo))
model.inla.ST2_poisson$cpo$failure

#DIC
model.inla.ST2_poisson$dic$dic
model.inla.ST2_poisson$dic$p.eff
model.inla.ST2_poisson$dic$mean.deviance
```

B.2.2 Negative Binomial

```
formula.ST2_nb<- y ~ empresas+setor_primario+setor_secundario+
   ↪ PIB_trim+iefp +sa6+sa8+
  f(ID.area,model="besag",graph=portugal.adj) +
  f(ID.year,model="rw1")+f(ID.area.year,model="iid")


model.inla.ST2_nb<- inla(formula.ST2_nb,family="nbinomial",data=
   ↪ data_paper, offset=log(E),control.predictor=list(compute=TRUE)
   ↪ , control.compute=list(waic=TRUE,dic=TRUE,cpo=TRUE))
```

B.2.3 Binomial

```
formula.ST2<- y ~ empresas+setor_primario+setor_secundario+PIB_trim+
   ↪ iefp +sa6+sa8+
  f(ID.area,model="besag",graph=portugal.adj) +
  f(ID.year,model="rw1")+f(ID.area.year,model="iid")

model.inla.ST2 <- inla(formula.ST2,family="binomial",data=data_paper,
   ↪ Ntrials=E, control.predictor=list(compute=TRUE), control.
   ↪ compute=list(waic=TRUE,dic=TRUE,cpo=TRUE))
```

B.2.4 Beta

```
formula.ST2_beta<- taxa ~ 1+empresas+setor_primario+setor_secundario+
    ↪ PIB_trim+iefp+sa6+sa8+
  f(ID.area,model="besag",graph=portugal.adj) +
  f(ID.year,model="rw1")+f(ID.area.year,model="iid")

model.inla.ST2_beta<- inla(formula.ST2_beta,family="beta",data=
    ↪ data_paper, control.predictor=list(compute=TRUE), control.
    ↪ compute=list(waic=TRUE,dic=TRUE,cpo=TRUE))
```

B.3 Spatial point processes models

B.3.1 Data preparation

```
#data from the 4th quarter of 2014

trim<-"4t14"

p<-paste(getwd(),"/dados/IE_pais_",trim,".csv",sep="")
p

base_dados<-read.table(paste(p),head=T,sep=",",dec=".",fileEncoding="
    ↪ latin1")
base_dados[1:3,]
dim(base_dados)

pesos<-read.table(paste(getwd(),"/dados/pesos",trim,".txt",sep=""),
    ↪ head=TRUE,fileEncoding="latin1")
head(pesos)
dim(pesos)

dados_1t<-merge(base_dados,pesos,by="IDENTIF_FIXA_COMPLETA")
dim(dados_1t)

#remove islands

dados_1t<-dados_1t[which(dados_1t$NUTS.3.NG!=200&dados_1t$NUTS.3.NG
    ↪ !=300),]

#binary variable: 1-unemployed, 0-c.c.

cpt2<-ifelse(dados_1t$CPT_PA==2,1,0)
```

```
dados_1t$cpt2<-cpt2


#identify each dwelling through its code

area_aloj<-paste(dados_1t$AREA_AM_ORIG,dados_1t$N_ALOJ_AM)

dados_1t$area_aloj<-area_aloj


#file with coordinates

coord4t14<-read.table(paste(getwd(),"/dados/coord_",trim,".txt",sep
    ↪ =""),head=TRUE,fileEncoding="latin1")
coord<-coord4t14

area_aloj<-paste(coord$AREA_AM,coord$AREA_AM_ALOJ)

coord$area_aloj<-area_aloj

#specify the coordinates of each dwelling

dados_finais<-merge(dados_1t,coord,by="area_aloj")
dim(dados_finais)

#remove duplicated points in the same dwelling

dados_finais2<-dados_finais[!duplicated(dados_finais$area_aloj),]
dim(dados_finais2)

#remove duplicated points in the same residential building

dados_finais5<-dados_finais2[!duplicated(dados_finais2$EDIF_COD),]
dim(dados_finais5)

#coordinates of the points

dados_marcas<-dados_finais5

coords<-dados_marcas[,c(27,28)]
coords4<-as.matrix(coords)

#transform the system of coordinates

library(sp)
coords_data <- SpatialPoints(coords4[,1:2])
```

```
library(rgdal)
d <- data.frame(lon=coords4[,1], lat=coords4[,2])
coordinates(d) <- c("lon", "lat")
proj4string(d) <- CRS("+init=epsg:3763") # WGS 84
CRS.new <- CRS("+init=epsg:4326 +proj=utm +zone=29 +units=km +ellps=
    ↪ WGS84 +datum=WGS84 +no_defs +towgs84=0,0,0")

d.ch1903 <- spTransform(d, CRS.new)

coords4<-coordinates(d.ch1903)

coords4<-as.matrix(coords4)

coords_in_4t14<-coords4


coords_in<-as.matrix(coords_in_4t14)
xy.cov<-coords_in

n.cov<-length(xy.cov[,1])
```

B.3.2 Mesh construction

```
#shapefile with mainland Portugal

cont<-readRDS("cont.rds")

#simplify the shapefile

library(rgeos)

gg<-gSimplify(cont,tol=0.5)


#construction of mesh

library(maptools)
nc.border <- unionSpatialPolygons(gg, rep(1, nrow(cont)))

library(INLA)
nc.bdry <- inla.sp2segment(nc.border)
```

```
(mesh <- inla.mesh.2d(boundary=nc.bdry,max.edge=c(10, 50),cutoff=0.1,
   ↪ offset=c(20,20)))$n



#PC prior for the mesh parameters

domain.size = 400
spde = inla.spde2.pcmatern(mesh,
                           prior.range=c(domain.size, 0.5),
                           prior.sigma=c(1, 0.5))
```

B.3.3 Covariates at observations and mesh nodes locations

```
#population density

raster_pop<-readRDS("raster_pop.rds")

est_dens_int<-extract(raster_pop,as.matrix(mesh$loc[,1:2]))
est_dens_loc<-extract(raster_pop,as.matrix(coords_in))

densidade_pop<-c(est_dens_int,est_dens_loc)



#coordinates for modelling the marks

coords_marcas<-rbind(mesh$loc[,1:2],coords_in)



#number of unemployed people per residential building

tt<-aggregate( dados_finais$cpt2 ~ dados_finais$EDIF_COD, FUN = sum )

names(tt)<-c("EDIF_COD","cpt3")


#number of dwellings per residential building

nh<-aggregate( dados_finais2$area_aloj ~ dados_finais2$EDIF_COD, FUN
   ↪ = length )

names(nh)<-c("EDIF_COD","naloj")
```

```
dados_marcas<-merge(dados_finais5,tt,by="EDIF_COD",sort=FALSE)

dados_marcas2<-merge(dados_marcas,nh,by="EDIF_COD",sort=FALSE)

y_4t14<-dados_marcas2$cpt3


#number of people per residential building

ttnind<-aggregate( dados_finais$cpt2 ~ dados_finais$EDIF_COD, FUN =
    ↪ length )

names(ttnind)<-c("EDIF_COD","cpt4")

dados_marcas4<-merge(dados_marcas,ttnind,by="EDIF_COD",sort=FALSE)

dados_marcas5<-merge(dados_marcas4,nh,by="EDIF_COD",sort=FALSE)

nind_offset<-dados_marcas5$cpt4


#mean age per residential building

idade<-aggregate( dados_finais$IDADE_ANOS ~ dados_finais$EDIF_COD,
    ↪ FUN = mean )

names(idade)<-c("EDIF_COD","idade")


#median of education level per residential building

edu<-aggregate( dados_finais$INSTRUCAO_PUB ~ dados_finais$EDIF_COD,
    ↪ FUN = median )

names(edu)<-c("EDIF_COD","edu")


dados_marcas5<-merge(dados_marcas4,idade,by="EDIF_COD",sort=FALSE)
head(dados_marcas5)

dados_marcas6<-merge(dados_marcas5,edu,by="EDIF_COD",sort=FALSE)
head(dados_marcas6)
```

```
#predict the covariates values at mesh nodes

library(INLA)
library(sp)
library(maptools)
library(rgeos)
library(spatstat)

cont<-readRDS("cont.rds")
gg<-gSimplify(cont,tol=0.5)
owin <- as.owin(gg)

pp_mark <- ppp(coords_in_4t14[,1],coords_in_4t14[,2],marks=
    ↪ nind_offset,window=owin)

kernel_marcas2 <- Smooth.ppp(pp_mark,sigma=20,edge = TRUE, diggle=
    ↪ TRUE)

raster_nind<-raster(kernel_marcas2)


nind<-extract(raster_nind,as.matrix(coords_in_4t14),method="bilinear
    ↪ ")

nind[is.na(nind)]<-mean(nind[!is.na(nind)])


nind_4t14<-nind

nind.v<-extract(raster_nind,as.matrix(mesh$loc[,1:2]),method="
    ↪ bilinear")

nind.v[is.na(nind.v)]<-mean(nind.v[!is.na(nind.v)])

nind.v_4t14<-nind.v



pp_mark <- ppp(coords_in_4t14[,1],coords_in_4t14[,2],marks=
    ↪ dados_marcas5$idade,window=owin)

kernel_marcas2 <- Smooth.ppp(pp_mark,sigma=5,edge = TRUE, diggle=TRUE
    ↪ )

raster_idade<-raster(kernel_marcas2)
```

```
covidade<-extract(raster_idade,as.matrix(coords_in_4t14),method="
    ↪ bilinear")

covidade[is.na(covidade)]<-mean(covidade[!is.na(covidade)])

covidade<-(covidade-mean(covidade))/sd(covidade)


covidade.v<-extract(raster_idade,as.matrix(mesh$loc[,1:2]),method="
    ↪ bilinear")

covidade.v2<-covidade.v
covidade.v2[is.na(covidade.v2)]<-mean(covidade.v2[!is.na(covidade.v2)
    ↪ ])
covidade.v3<-(covidade.v2-mean(covidade.v2))/sd(covidade.v2)


pp_mark <- ppp(coords_in_4t14[,1],coords_in_4t14[,2],marks=
    ↪ dados_marcas6$edu,window=owin)

kernel_marcas2 <- Smooth.ppp(pp_mark,sigma=5,edge = TRUE, diggle=TRUE
    ↪ )

raster_edu<-raster(kernel_marcas2)


covedu<-extract(raster_edu,as.matrix(coords_in_4t14),method="bilinear
    ↪ ")

covedu[is.na(covedu)]<-mean(covedu[!is.na(covedu)])
covedu<-(covedu-mean(covedu))/sd(covedu)

covedu.v<-extract(raster_edu,as.matrix(mesh$loc[,1:2]),method="
    ↪ bilinear")

covedu.v[is.na(covedu.v)]<-mean(covedu.v[!is.na(covedu.v)])

covedu.v<-(covedu.v-mean(covedu.v))/sd(covedu.v)


covedu_4t14<-covedu
covedu.v_4t14<-covedu.v

covidade_4t14<-covidade
```

```
covidade.v3_4t14<-covidade.v3


#shapefile with municipalities in mainland Portugal

concelhos<-readRDS("concelhos.rds")


#population estimates by municipality for 2014

pop_concelho<-read.table(paste(getwd(),"/dados/pop_concelho2014.txt",
   ↪ sep=""))


names(pop_concelho)<-c("codigo","pop")


library(stringr)
pop_concelho$dico<-str_sub(pop_concelho$codigo, start= -4)


#data from the IEFP by municipality for the 4th quarter of 2014

iefp_fev<-read.table(paste(getwd(),"/dados/iefp_nov14.txt",sep=""),
   ↪ colClasses=c("character","numeric"))


names(iefp_fev)<-c("dico","iefp")

iefp_fev2<-merge(pop_concelho,iefp_fev,by="dico")

iefp_fev2$prop_iefp<-iefp_fev2$iefp/iefp_fev2$pop


#predict the IEFP covariate at mesh nodes

iefp_ppp2<-ppp(coord_centroides[,1], coord_centroides[,2],window=owin
   ↪ ,marks=iefp_fev2$prop_iefp)


kernel_iefp2 <- Smooth.ppp(iefp_ppp2,sigma=20,edge = TRUE, diggle=
   ↪ TRUE)


library(raster)
raster_iefp<-raster(kernel_iefp2)
```

```
raster_iefp_4t14<-raster_iefp

iefp<-extract(raster_iefp_4t14,as.matrix(coords_in_4t14),method="
   ↪ bilinear")

iefp[is.na(iefp)]<-mean(iefp[!is.na(iefp)])
iefp<-(iefp-mean(iefp))/sd(iefp)

iefp.v<-extract(raster_iefp_4t14,as.matrix(mesh$loc[,1:2]),method="
   ↪ bilinear")

iefp.v[is.na(iefp.v)]<-mean(iefp.v[!is.na(iefp.v)])

iefp.v<-(iefp.v-mean(iefp.v))/sd(iefp.v)


iefp_4t14<-iefp
iefp.v_4t14<-iefp.v


marcas<-c(rep(NA,nv),y_4t14)

est_nind_4t14<-c(nind.v_4t14,nind_4t14)
est_covedu_4t14<-c(covedu.v_4t14,covedu_4t14)
est_covidade_4t14<-c(covidade.v3_4t14,covidade_4t14)
est_iefp_4t14<-c(iefp.v_4t14,iefp_4t14)
```

B.3.4 inla.stack syntax

```
#number of mesh nodes

nv<-mesh$n

#triangulation weights

w<-diag(inla.mesh.fem(mesh)$c0)


#projection matrix

lmat.cov<-inla.spde.make.A(mesh,xy.cov)
```

```
imat <- Diagonal(nv, rep(1, nv))

A.pp.cov<-rBind(imat,lmat.cov)


#coordinates for modelling the marks

coords_marcas<-rbind(mesh$loc[,1:2],coords_in)


#projection matrix

lmat.cov.marcas<-inla.spde.make.A(mesh,coords_marcas)


#data for modelling the points

y.pp.cov<-rep(0:1,c(nv,n.cov))
e.pp.cov<-c(w,rep(0,n.cov))


stk.pp <- inla.stack(data=list(y=cbind(y.pp.cov,NA), e=e.pp.cov,link
   ↪ =1), A=list(1,A.pp.cov), tag='pp',
                  effects=list(list(b0.pp=rep(1,nv+n.cov),dens.pp=
                     ↪ densidade_pop), list(i=1:nv,iidx=1:nv)))



#data for modelling the marks



stk.m2 <- inla.stack(data=list(y=cbind(NA,marcas),e=c(rep(1,nv),rep
   ↪ (1,n.cov)),link=1), A=list(lmat.cov.marcas, 1), tag='marcas',
                  effects=list(list(j=1:nv,k=1:nv,iidx2=1:nv),
                  list(b0.y=rep(1,nv+n.cov),
                   Ntrial=est_nind_4t14,
                   cov.edu=est_covedu_4t14,
                   cov.idade=est_covidade_4t14,
                   iefp=est_iefp_4t14
                  )))


#data for points and marks

j.stk <- inla.stack(stk.pp, stk.m2)
```

B.3.5 Marked LGCP model

```
#marked LGCP model

jform0011 <- y ~ -1 + b0.pp + b0.y + offset(log(dens.pp)) + Ntrial +
    ↪ cov.edu + cov.idade + iefp + f(i, model=spde)+
  f(k, model=spde)


j.res0011 <- inla(jform0011, family=c('poisson', 'poisson'), data=
    ↪ inla.stack.data(j.stk), E=inla.stack.data(j.stk)$e,
              control.predictor=list(A=inla.stack.A(j.stk),compute=
                 ↪ TRUE,link=1),
              control.compute=list(config=TRUE,dic=TRUE,cpo=TRUE,waic
                 ↪ =TRUE),
              quantiles=c(0.025, 0.5, 0.975),
              control.results=list(return.marginals.random=F,return.
                 ↪ marginals.predictor=F),
              control.inla=list(strategy="gaussian"),
              inla.call="remote")
```

B.3.6 Estimates at NUTS III level

```
#index for marks

idx.marcas<-inla.stack.index(j.stk, tag="marcas")$data
idx.marcas.v<-idx.marcas[1]:(idx.marcas[1]+nv-1)

#index for points

idx.pp<-inla.stack.index(j.stk, tag="pp")$data
idx.pp.v<-idx.pp[1]:(idx.pp[1]+nv-1)


#predicted values of intensity of points at mesh nodes

fitted<-j.res0011$summary.fitted.values[idx.pp.v,1]


#projection from the mesh to the grid

(nxy <- round(c(diff(c(455.489, 734.3417)), diff(c(4091.206
    ↪ ,4667.2201)))))
```

```
projgrid <- inla.mesh.projector(mesh, xlim=c(455.489, 734.3417),
                                ylim=c(4091.206 ,4667.2201), dims=nxy)

cellArea<-diff(projgrid$x)*diff(projgrid$y)


xmean_cov <- inla.mesh.project(projgrid, fitted)

xmean_cov_marcas <- inla.mesh.project(projgrid, j.res0011$summary.
    ↪ fitted.values[idx.marcas.v,1])


#selection probabilities of the dwellings

dados_finais2<-readRDS("dados_finais2.rds")

owin <- as.owin(gg)

library(spatstat)

pp_mark <- ppp(coords_in[,1],coords_in[,2],marks=dados_finais2$prob,
    ↪ window=owin)

kernel_prob20 <- Smooth.ppp(pp_mark,sigma=20,edge = TRUE, diggle=TRUE
    ↪ )

library(raster)

raster_prob<-raster(kernel_prob20)

est_prob_grid<-extract(raster_prob,as.matrix(projgrid$lattice$loc
    ↪ [,1:2]),method="bilinear")

est_prob_int<-extract(raster_prob,as.matrix(mesh$loc[,1:2]),method="
    ↪ bilinear")


#estimates for the total unemployed for each grid cell

xmean_est_area<-xmean_cov*xmean_cov_marcas*cellArea/est_prob_grid


#shapefile with NUTS III regions

nuts3<-readRDS("nuts3.rds")
```

```
coords_data <- SpatialPoints(projgrid$lattice$loc)
proj4string(coords_data)<-proj4string(nuts3)


#grid cells over NUTS III regions

match_coords_asl <- over(coords_data,nuts3)

match_coords_asl$xmean_s<-c(xmean_est_area)

#estimates for the total unemployed by NUTS III

est_nuts3_s<-aggregate( match_coords_asl$xmean_s ~
   ↪ match_coords_asl$NUTS3_02, FUN = sum )

names(est_nuts3_s)<-c("NUTS3_02","est")


#1000 samples from the model

samples_m = inla.posterior.sample(1000,j.res0011)


#estimates for the variances by NUTS III

nsamples<-1000

est_nuts3_amostras<-est_nuts3_s
match_coords_asl$xmean<-0


for (k in 1:nsamples){

fitted<-exp(samples_m[[k]]$latent[idx.pp.v,])

xmean_cov <- inla.mesh.project(projgrid, fitted)

xmean_cov_marcas <- inla.mesh.project(projgrid, exp(samples_m[[k]]
   ↪ $latent[idx.marcas.v,]))

xmean_est_area<-xmean_cov*xmean_cov_marcas*cellArea*est_prob_grid

match_coords_asl$xmean_s<-c(xmean_est_area)

est_nuts3_s<-aggregate( match_coords_asl$xmean_s ~
   ↪ match_coords_asl$NUTS3_02, FUN = sum )
```

```
names(est_nuts3_s)<-c("NUTS3_02","est")

est_nuts3_amostras<-cbind(est_nuts3_amostras,est_nuts3_s$est)

}


library(matrixStats)

medias<-rowMeans(est_nuts3_amostras[,3:nsamples+2])
variancias<-rowVars(as.matrix(est_nuts3_amostras[,3:nsamples+2]))
cv<-sqrt(variancias)/medias
```

B.4 Geostatistics models

B.4.1 Mesh construction

```
library(INLA)
library(raster)

#coordinates of residential buildings in the LFS from the 4th quarter
    ↪  of 2014 to the 4th quarter of 2016

dat<-readRDS("dat4.rds")

coords_in<-as.matrix(data.frame(dat$xcoo,dat$ycoo))

prdomain <- inla.nonconvex.hull(coords_in, -0.03, -0.05, resolution=c
    ↪ (100,100))

#mesh construction

prmesh <- inla.mesh.2d(boundary=prdomain, max.edge=c(25,50), cutoff
    ↪ =2)

mesh<-prmesh


#PC prior for the SPDE parameters

domain.size = 400
spde = inla.spde2.pcmatern(mesh,
                           prior.range=c(domain.size, 0.5),
                           prior.sigma=c(1, 0.5))
```

B.4.2 inla.stack syntax

```
#define the quarter for the prediction: 4t16

mesh.index<-inla.spde.make.index(name="field",n.spde=spde$n.spde,n.
   ↪ group=9)


#projection matrix

A.est <-inla.spde.make.A(mesh=mesh,
                         loc=cbind(dat$xcoo, dat$ycoo),
                         group=dat$time,
                         n.group=9)

#data at observations locations

stack.est <- inla.stack(data=list(y=dat$y), A=list(A.est,1), tag='
   ↪ marcas',
                        effects=list(c(mesh.index,list(b0.y=1)),
                        list(Ntrial=nindc,
                          cov.edu=coveduc,
                          cov.idade=covidadec,
                          iefp=iefpc
                        )))

#projection matrix

A.pred<-inla.spde.make.A(mesh,group=9,n.group=9)


#data at mesh nodes locations

stack.pred<- inla.stack(data=list(y=NA),A=list(A.pred,1),tag="pred",
                        effects=list(c(mesh.index,list(b0.y=1)),
                        list(Ntrial=nind.v_4t16,
                        cov.edu=covedu.v_4t16,
                        cov.idade=covidade.v3_4t16,
                        iefp=iefp.v_4t16)))

#joint data

join.stack<-inla.stack(stack.est,stack.pred)
```

B.4.3 Geostatistical data model

```
#Prior for the parameters of AR(1) model

h.spec <- list(theta=list(prior='pccor1', param=c(0, 0.9)))

#Geostatistical data model

formula6 <- y ~ -1 + b0.y + offset(log(Ntrial)) + cov.idade + iefp +
   ↪ f(field, model=spde,group=field.group,control.group=list(model
   ↪ ='ar1', hyper=h.spec))


r6_4t16_eb <- inla(formula6, family='poisson',data=inla.stack.data(
   ↪ join.stack,spde=spde),
              control.predictor=list(A=inla.stack.A(join.stack),
                 ↪ compute=TRUE,link=1),
              control.compute=list(config=TRUE,dic=TRUE,cpo=TRUE,
                 ↪ waic=TRUE),
              quantiles=c(0.025, 0.5, 0.975),
              control.results=list(return.marginals.random=F,return.
                 ↪ marginals.predictor=F),
              control.inla=list(strategy="gaussian",int.strategy="eb
                 ↪ "),
              inla.call="remote")


#use rerun to make the model more stable

inla.r6 = inla.rerun(r6_4t16_eb)
```

B.4.4 Estimates at NUTS III level

```


#projection from the mesh to the grid

(nxy <- round(c(diff(c(455.489, 734.3417)), diff(c(4091.206
   ↪ ,4667.2201)))))

projgrid <- inla.mesh.projector(mesh, xlim=c(455.489, 734.3417),
                                ylim=c(4091.206 ,4667.2201), dims=nxy)
```

```
#number of dwellings in each grid cell

table5<-readRDS("table5.rds")


#define as NA outside the domain of interest (mainland Portugal)

raster_prob<-readRDS("raster_prob.rds")

est_prob_grid<-extract(raster_prob,as.matrix(projgrid$lattice$loc
   ↪ [,1:2]),method="bilinear")

est_prob_grid2<-est_prob_grid
est_prob_grid2[!is.na(est_prob_grid2)]<-1


#index for predictions

idx.marcas1<-inla.stack.index(join.stack, tag="pred")$data


#estimates at grid cell level

xmean_cov_marcas <- inla.mesh.project(projgrid, inla.r6$summary.
   ↪ fitted.values[idx.marcas1,1])

xmean_est_area<-xmean_cov_marcas*table5*est_prob_grid2

total_desemp4t16<-xmean_est_area


#estimates at NUTS III level


nuts3_2013<-readRDS("nuts3_2013.rds")

coords_data <- SpatialPoints(projgrid$lattice$loc)
proj4string(coords_data)<-proj4string(nuts3_2013)

match_coords_asl <- over(coords_data,nuts3_2013)


match_coords_asl$xmean_s<-c(xmean_est_area)
```

```
est_nuts3_s<-aggregate( match_coords_asl$xmean_s ~
   ↪ match_coords_asl$NUTS3_15DE, FUN = sum )


names(est_nuts3_s)<-c("NUTS3_2013","est")

totais_4t16_v2013<-est_nuts3_s$est


#variances


samples_m = inla.posterior.sample(1000,inla.r6)

idx.marcas1<-inla.stack.index(join.stack, tag="pred")$data

nsamples<-1000

est_nuts3_p<-0

for (k in 1:nsamples){


  xmean_cov_marcas <- inla.mesh.project(projgrid, exp(samples_m[[k]]
     ↪ $latent[idx.marcas1,]))


  xmean_est_area<-xmean_cov_marcas*table5*est_prob_grid2

  match_coords_asl$xmean_s<-c(xmean_est_area)

  est_nuts3_s<-aggregate( match_coords_asl$xmean_s ~
     ↪ match_coords_asl$NUTS3_15DE, FUN = sum )
  names(est_nuts3_s)<-c("NUTS3_2013","est")


  est_nuts3_p<-cbind(est_nuts3_p,est_nuts3_s$est)

}

library(matrixStats)

medias_g<-rowMeans(est_nuts3_p[,5:(nsamples)])
variancias_g<-rowVars(as.matrix(est_nuts3_p[,5:(nsamples)]))
cv_g<-sqrt(variancias_g)/medias_g
```

B.5 Maps

```
#maps at grid cell level

est_prob_grid2<-est_prob_grid
est_prob_grid2[!is.na(est_prob_grid2)]=1

xmean_est_area1<-xmean_cov*est_prob_grid2
xmean_est_area2<-xmean_cov_marcas*est_prob_grid2


library(gridExtra)
library(lattice)
library(fields)

grid.arrange(levelplot(xmean_est_area1, col.regions=tim.colors(99),
                       xlab='', ylab='', scales=list(draw=FALSE)),
             levelplot(xmean_est_area2, col.regions=tim.colors(99),
                       xlab='', ylab='', scales=list(draw=FALSE)),ncol
                          ↪ =2)



#maps at NUTS III level

library(sp)
library(maps)
library(maptools)
library(rgdal)
library(ggplot2)
library(plyr)
library(broom)
library(rgeos)

#shapefile with the NUTS III regions

nuts3<-readRDS("nuts3_2013.rds")

#simplify the shapefile

nuts3_2<-gSimplify(nuts3,tol=0.5)
nuts3_dt<-tidy(nuts3_2)

#names of NUTS III regions
```

```
temp_df <- data.frame(nuts3@data$NUTS3_15DE)
names(temp_df) <- c("NUTS3_15DE")

# create and append "id"

temp_df$id <- seq(0,nrow(temp_df)-1)
new_df <- join(nuts3_dt, temp_df, by="id")


#data (for 4 maps)

td1<-cv_dir
td2<-cv
td3<-cv_p
td4<-cv_g

indice_nuts3<-readRDS("est_nuts3_s$NUTS3_2013.rds")

data<-data.frame(indice_nuts3,td1,td2,td3,td4)
names(data)<-c("NUTS3_15DE","td1","td2","td3","td4")


#merge data and shapefile

county.df<-merge(new_df,data, by=intersect(names(new_df), names(data)
    ↪ ))

td1<-data.frame(county.df$td1,county.df$long,county.df$lat,county.
    ↪ df$group)
names(td1)<-c("county.df.td", "county.df.long", "county.df.lat" , "
    ↪ county.df.group")

td1$id = 'direct'

td1$id2 = 1


td2<-data.frame(county.df$td2,county.df$long,county.df$lat,county.
    ↪ df$group)
names(td2)<-c("county.df.td", "county.df.long", "county.df.lat" , "
    ↪ county.df.group")

td2$id = 'areal'

td2$id2 = 2
```

```
td3<-data.frame(county.df$td3,county.df$long,county.df$lat,county.
    ↪ df$group)
names(td3)<-c("county.df.td", "county.df.long", "county.df.lat" , "
    ↪ county.df.group")

td3$id = 'lgcp'

td3$id2 = 3

td4<-data.frame(county.df$td4,county.df$long,county.df$lat,county.
    ↪ df$group)
names(td4)<-c("county.df.td", "county.df.long", "county.df.lat" , "
    ↪ county.df.group")

td4$id = 'geost'

td4$id2 = 4

df_all = rbind(td1,td2,td3,td4)

df_all[1:30,]
names(df_all)<-c("Y","longitude","latitude","grupo","id","id2")

df_all$id<-factor(df_all$id,levels=c("direct","areal","lgcp","geost")
    ↪ )


#maps

library(fields)
ggplot(data=df_all, aes(x=longitude, y=latitude, group=grupo)) +
  geom_polygon(aes(fill=Y)) +
  coord_equal() +
  scale_fill_gradientn(colours=tim.colors(99))+
  facet_wrap(~id,ncol=4)
```